

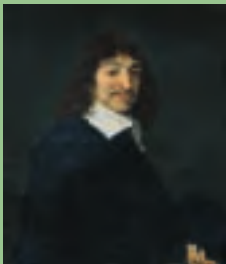


9

- 9.1 Scatter Diagrams and Linear Correlation
- 9.2 Linear Regression and the Coefficient of Determination
- 9.3 Inferences for Correlation and Regression
- 9.4 Multiple Regression



Asia Images Group/Getty Images



Portrait of René Descartes c. 1649 (after Frans Hals/Louvre/Giraudon/The Bridgeman Art Library)

When it is not in our power to determine what is true, we ought to follow what is most probable.

—RENÉ DESCARTES

It is important to realize that statistics and probability do not deal in the realm of certainty. If there is any realm of human knowledge where genuine certainty exists, you may be sure that our statistical methods are not needed there. In most human endeavors, and in almost all of the natural world around us, the element of chance happenings cannot be avoided. When we cannot expect something with true certainty, we must rely on probability to be our guide. In this chapter, we will study regression, correlation, and forecasting. One of the tools we use is a scatter plot. René Decartes (1596–1650) was the first mathematician to systematically use rectangular coordinate plots. For this reason, such a coordinate axis is called a Cartesian axis.

For online student resources, visit the Brase/Brase, *Understandable Statistics*, 10th edition web site at <http://www.cengage.com/statistics/brase>

CORRELATION AND REGRESSION

PREVIEW QUESTIONS

How can you use a scatter diagram to visually estimate the degree of linear correlation of two random variables? (SECTION 9.1)

How do you compute the correlation coefficient and what does it tell you about the strength of the linear relationship between two random variables? (SECTION 9.1)

What is the least-squares criterion? How do you find the equation of the least-squares line? (SECTION 9.2)

What is the coefficient of determination and what does it tell you about explained variation of y in a random sample of data pairs (x, y) ? (SECTION 9.2)

How do you determine if the sample correlation coefficient is statistically significant? (SECTION 9.3)

How do you find a confidence interval for predictions based on the least-squares model? (SECTION 9.3)

How do you test the slope β of the population least-squares line? How do you construct a confidence interval for β ? (SECTION 9.3)

What if you have more than two random variables? How do you construct a linear regression model for three, four, or more random variables? (SECTION 9.4)

David Young-Wolff/PhotoEdit



FOCUS PROBLEM

Changing Populations and Crime Rate

Is the crime rate higher in neighborhoods where people might not know each other very well? Is there a relationship between crime rate and population change? If so, can we make predictions based on such a relationship? Is the relationship statistically significant? Is it possible to predict crime rates from population changes?

Denver is a city that has had a lot of growth and consequently a lot of population change in recent years. Sociologists studying population changes and crime rates could find a wealth of information in Denver statistics. Let x be a random variable representing percentage change in neighborhood population in the past few years, and let y be a random variable representing crime rate (crimes per 1000 population). A random sample of

ofoto, 2010/Used under license from Shutterstock.com



six Denver neighborhoods gave the following information (Source: *Neighborhood Facts*, The Piton Foundation). To find out more about the Piton Foundation, visit the Brase/Brase statistics site at <http://www.cengage.com/statistics/brase> and find the link to the Piton Foundation.

x	29	2	11	17	7	6
y	173	35	132	127	69	53

Using information presented in this chapter, you will be able to analyze the relationship between the variables x and y using the following tools.

- Scatter diagram
- Sample correlation coefficient and coefficient of determination
- Least-squares line equation
- Predictions for y using the least-squares line
- Tests of population correlation coefficient and of slope of least-squares line
- Confidence intervals for slope and for predictions

(See Problem 10 in the Chapter Review Problems.)

SECTION 9.1

Scatter Diagrams and Linear Correlation

FOCUS POINTS

- Make a scatter diagram.
- Visually estimate the location of the “best-fitting” line for a scatter diagram.
- Use sample data to compute the sample correlation coefficient r .
- Investigate the meaning of the correlation coefficient r .

Paired data values

Scatter diagram

Explanatory variable
Response variable

Studies of correlation and regression of two variables usually begin with a graph of *paired data values* (x, y) . We call such a graph a *scatter diagram*.

A **scatter diagram** is a graph in which data pairs (x, y) are plotted as individual points on a grid with horizontal axis x and vertical axis y . We call x the **explanatory variable** and y the **response variable**.

By looking at a scatter diagram of data pairs, you can observe whether there seems to be a linear relationship between the x and y values.

EXAMPLE 1

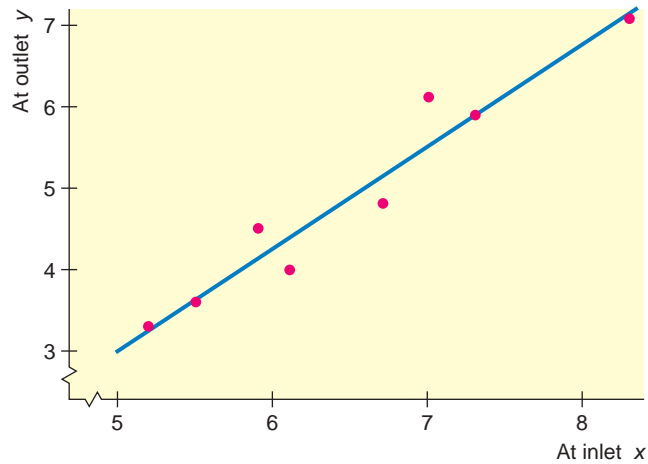
SCATTER DIAGRAM

Phosphorous is a chemical used in many household and industrial cleaning compounds. Unfortunately, phosphorous tends to find its way into surface water, where it can kill fish, plants, and other wetland creatures. Phosphorous-reduction programs are required by law and are monitored by the Environmental Protection Agency (EPA) (Reference: *EPA Case Study 832-R-93-005*).

A random sample of eight sites in a California wetlands study gave the following information about phosphorous reduction in drainage water. In this study, x is a random variable that represents phosphorous concentration (in 100 mg/l) at the inlet of a passive biotreatment facility, and y is a random

FIGURE 9-1

Phosphorous Reduction (100 mg/l)



variable that represents total phosphorous concentration (in 100 mg/l) at the outlet of the passive biotreatment facility.

x	5.2	7.3	6.7	5.9	6.1	8.3	5.5	7.0
y	3.3	5.9	4.8	4.5	4.0	7.1	3.6	6.1

(a) Make a scatter diagram for these data.

SOLUTION: Figure 9-1 shows points corresponding to the given data pairs. These plotted points constitute the scatter diagram. To make the diagram, first scan the data and decide on an appropriate scale for each axis. Figure 9-1 shows the scatter diagram (points) along with a line segment showing the basic trend. Notice a “jump scale” on both axes.

(b) **Interpretation** Comment on the relationship between x and y shown in Figure 9-1.

SOLUTION: By inspecting the figure, we see that smaller values of x are associated with smaller values of y and larger values of x tend to be associated with larger values of y . Roughly speaking, the general trend seems to be reasonably well represented by an upward-sloping line segment, as shown in the diagram.



Cindy Kassab/Corbis Edge/Corbis

Of course, it is possible to draw many curves close to the points in Figure 9-1, but a straight line is the simplest and most widely used in elementary studies of paired data. We can draw many lines in Figure 9-1, but in some sense, the “best” line should be the one that comes closest to each of the points of the scatter diagram. To single out one line as the “best-fitting line,” we must find a mathematical criterion for this line and a formula representing the line. This will be done in Section 9.2 using the *method of least squares*.

Another problem precedes that of finding the “best-fitting line.” That is the problem of determining how well the points of the scatter diagram are suited for fitting *any* line. Certainly, if the points are a very poor fit to any line, there is little use in trying to find the “best” line.

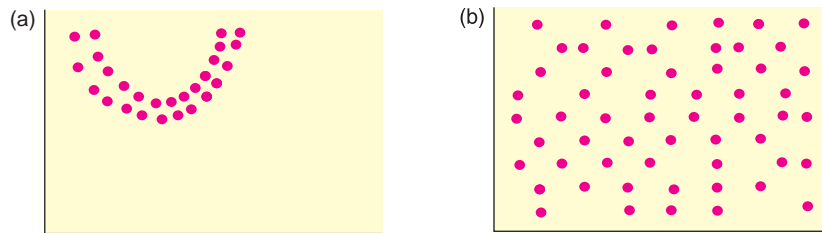
If the points of a scatter diagram are located so that *no* line is realistically a “good” fit, we then say that the points possess *no linear correlation*. We see some examples of scatter diagrams for which there is no linear correlation in Figure 9-2.

Introduction to linear correlation

No linear correlation

FIGURE 9-2

Scatter Diagrams with No Linear Correlation



GUIDED EXERCISE 1

Scatter diagram

A large industrial plant has seven divisions that do the same type of work. A safety inspector visits each division of 20 workers quarterly. The number x of work-hours devoted to safety training and the number y of work-hours lost due to industry-related accidents are recorded for each separate division in Table 9-1.



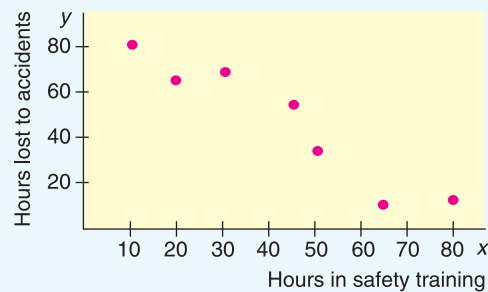
TABLE 9-1 Safety Report

Division	x	y
1	10.0	80
2	19.5	65
3	30.0	68
4	45.0	55
5	50.0	35
6	65.0	10
7	80.0	12

(a) Make a scatter diagram for these pairs. Place the x values on the horizontal axis and the y values on the vertical axis.



FIGURE 9-3 Scatter Diagram for Safety Report



(b) As the number of hours spent on safety training increases, what happens to the number of hours lost due to industry-related accidents?



In general, as the number of hours in safety training goes up, the number of hours lost due to accidents goes down.

(c) *Interpretation* Does a line fit the data reasonably well?



A line fits reasonably well.

(d) Draw a line that you think “fits best.”



Use a downward-sloping line that lies close to the points. Later, you will find the equation of the line that is a “best fit.”

If the points seem close to a straight line, we say the linear correlation is moderate to high, depending on how close the points lie to the line. If all the points do, in fact, lie on a line, then we have *perfect linear correlation*. In Figure 9-4, we see some diagrams with perfect linear correlation. In statistical applications, perfect linear correlation almost never occurs.

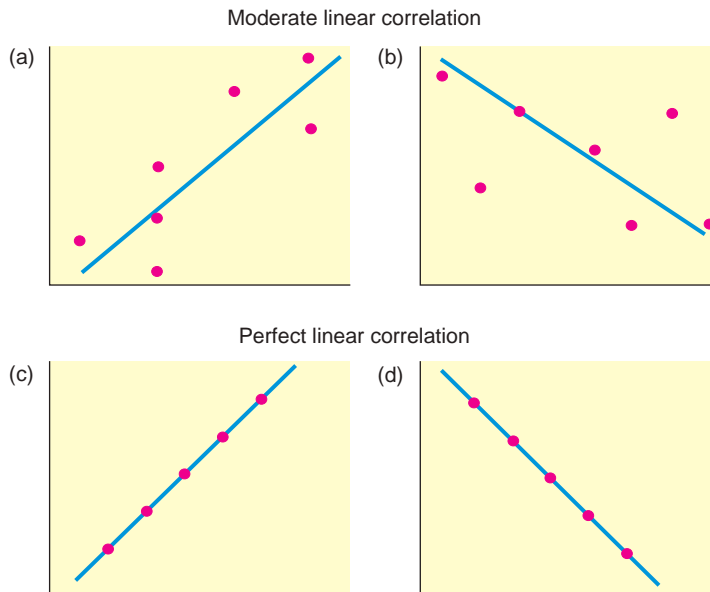
The variables x and y are said to have *positive correlation* if low values of x are associated with low values of y and high values of x are associated with high

Perfect linear correlation

Positive correlation

FIGURE 9-4

Scatter Diagrams with Moderate and Perfect Linear Correlation



Scatter diagrams for the same data look different from one another when they are graphed using different scales. Problem 19 at the end of this section explores how changing scales affect the look of a scatter diagram.

values of y . Figure 9-4 parts (a) and (c) show scatter diagrams in which the variables are positively correlated. On the other hand, if low values of x are associated with high values of y and high values of x are associated with low values of y , the variables are said to be *negatively correlated*. Figure 9-4 parts (b) and (d) show variables that are negatively correlated.

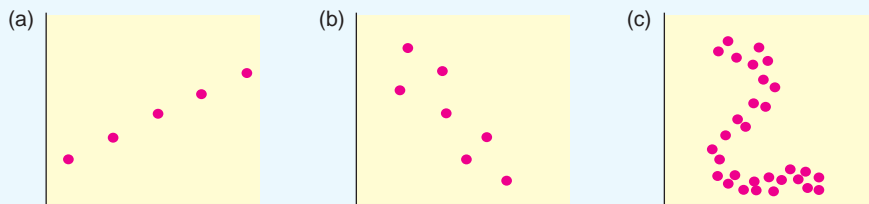
Negative correlation

GUIDED EXERCISE 2

Scatter diagram and linear correlation

Examine the scatter diagrams in Figure 9-5 and then answer the following questions.

FIGURE 9-5 Scatter Diagrams



- (a) Which diagram has no linear correlation? ➔ Figure 9-5(c) has no linear correlation. No straight-line fit should be attempted.
- (b) Which has perfect linear correlation? ➔ Figure 9-5(a) has perfect linear correlation and can be fitted exactly by a straight line.
- (c) Which can be reasonably fitted by a straight line? ➔ Figure 9-5(b) can be reasonably fitted by a straight line.



TECH NOTES

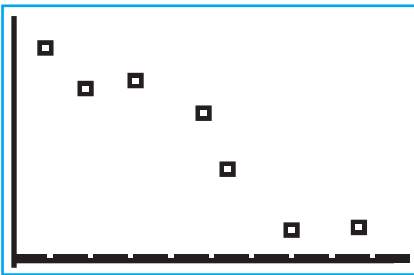
The TI-84Plus/TI-83Plus/TI-*n*spire calculators, Excel 2007, and Minitab all produce scatter plots. For each technology, enter the x values in one column and the corresponding y values in another column. The displays on the next page show the data

from Guided Exercise 1 regarding safety training and hours lost because of accidents. Notice that the scatter plots do not necessarily show the origin.

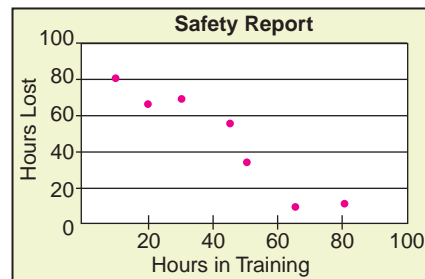
TI-84Plus/TI-83Plus/TI-*nspire* (with TI-84Plus keypad) Enter the data into two columns. Use **Stat Plot** and choose the first type. Use option **9: ZoomStat** under **Zoom**. To check the scale, look at the settings displayed under **Window**.

Excel 2007 Enter the data into two columns. On the home screen, click the **Insert** tab. In the Chart Group, select **Scatter** and choose the first type. In the next ribbon, the Chart Layout Group offers options for including titles and axes labels. Right clicking on data points provides other options such as data labels. Changing the size of the diagram box changes the scale on the axes.

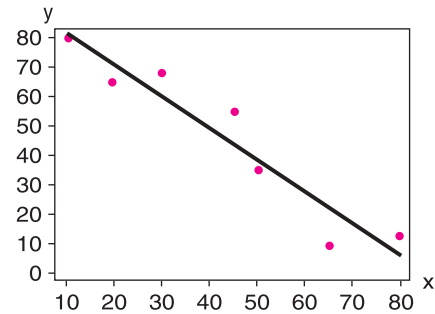
Minitab Enter the data into two columns. Use the menu selections **Stat** ► **Regression** ► **Fitted Line Plot**. The best-fit line is automatically plotted on the scatter diagram.

TI-84Plus/TI-83Plus/TI-*nspire* Display

Excel Display



Minitab Display



Sample Correlation Coefficient r

Looking at a scatter diagram to see whether a line best describes the relationship between the values of data pairs is useful. In fact, whenever you are looking for a relationship between two variables, making a scatter diagram is a good first step.

There is a mathematical measurement that describes the strength of the linear association between two variables. This measure is the *sample correlation coefficient* r . The full name for r is the *Pearson product-moment correlation coefficient*, named in honor of the English statistician Karl Pearson (1857–1936), who is credited with formulating r .

Sample correlation coefficient r

The **sample correlation coefficient** r is a numerical measurement that assesses the strength of a *linear* relationship between two variables x and y .

1. r is a unitless measurement between -1 and 1 . In symbols, $-1 \leq r \leq 1$. If $r = 1$, there is perfect positive linear correlation. If $r = -1$, there is perfect negative linear correlation. If $r = 0$, there is no linear correlation. The closer r is to 1 or -1 , the better a line describes the relationship between the two variables x and y .
2. Positive values of r imply that as x increases, y tends to increase. Negative values of r imply that as x increases, y tends to decrease.
3. The value of r is the same regardless of which variable is the explanatory variable and which is the response variable. In other words, the value of r is the same for the pairs (x, y) and the corresponding pairs (y, x) .
4. The value of r does not change when either variable is converted to different units.

We'll develop the defining formula for r and then give a more convenient computation formula.

Development of Formula for r

If there is a *positive* linear relation between variables x and y , then high values of x are paired with high values of y , and low values of x are paired with low values of y . [See Figure 9-6(a).] In the case of *negative* linear correlation, high values of x are paired with low values of y , and low values of x are paired with high values of y . This relation is pictured in Figure 9-6(b). If there is *little or no linear correlation* between x and y , however, then we will find both high and low x values sometimes paired with high y values and sometimes paired with low y values. This relation is shown in Figure 9-6(c).

These observations lead us to the development of the formula for the sample correlation coefficient r . Taking *high* to mean “above the mean,” we can express the relationships pictured in Figure 9-6 by considering the products

$$(x - \bar{x})(y - \bar{y})$$

If both x and y are high, both factors will be positive, and the product will be positive as well. The sign of this product will depend on the relative values of x and y compared with their respective means.

$$(x - \bar{x})(y - \bar{y}) \begin{cases} \text{is positive if } x \text{ and } y \text{ are both “high”} \\ \text{is positive if } x \text{ and } y \text{ are both “low”} \\ \text{is negative if } x \text{ is “low,” but } y \text{ is “high”} \\ \text{is negative if } x \text{ is “high,” but } y \text{ is “low”} \end{cases}$$

In the case of positive linear correlation, most of the products $(x - \bar{x})(y - \bar{y})$ will be positive, and so will the sum over all the data pairs

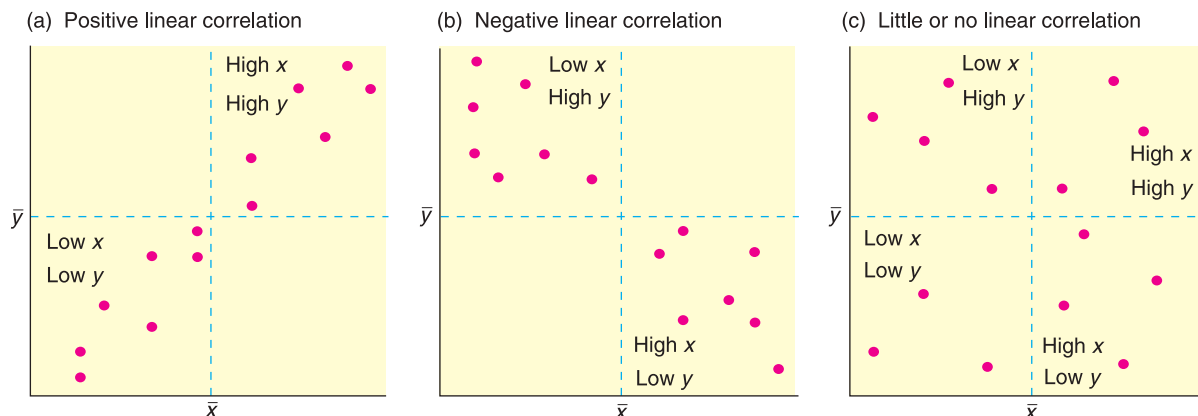
$$\Sigma(x - \bar{x})(y - \bar{y})$$

For negative linear correlation, the products will tend to be negative, so the sum also will be negative. On the other hand, in the case of little, if any, linear correlation, the sum will tend to be zero.

One trouble with the preceding sum is that it increases or decreases, depending on the units of x and y . Because we want r to be unitless, we standardize both x and y of a data pair by dividing each factor $(x - \bar{x})$ by the sample standard deviation s_x and each factor $(y - \bar{y})$ by s_y . Finally, we take an average of all the

FIGURE 9-6

Patterns for Linear Correlation



products. For technical reasons, we take the average by dividing by $n - 1$ instead of by n . This process leads us to the desired measurement, r .

$$r = \frac{1}{n - 1} \sum \frac{(x - \bar{x})}{s_x} \cdot \frac{(y - \bar{y})}{s_y} \quad (1)$$

Computation Formula for r

The defining formula for r shows how the mean and standard deviation of each variable in the data pair enter into the formulation of r . However, the defining formula is technically difficult to work with because of all the subtractions and products. A computation formula for r uses the raw data values of x and y directly.

PROCEDURE

HOW TO COMPUTE THE SAMPLE CORRELATION COEFFICIENT r

Requirements

Obtain a random sample of n data pairs (x, y) . The data pairs should have a *bivariate normal distribution*. This means that for a fixed value of x , the y values should have a normal distribution (or at least a mound-shaped and symmetric distribution), and for a fixed y , the x values should have their own (approximately) normal distribution.

Procedure

1. Using the data pairs, compute Σx , Σy , Σx^2 , Σy^2 , and Σxy .
2. With $n =$ sample size, Σx , Σy , Σx^2 , Σy^2 , and Σxy , you are ready to compute the sample correlation coefficient r using the computation formula

$$r = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{n\Sigma x^2 - (\Sigma x)^2} \sqrt{n\Sigma y^2 - (\Sigma y)^2}} \quad (2)$$

Be careful! The notation Σx^2 means first square x and then calculate the sum, whereas $(\Sigma x)^2$ means first sum the x values and then square the result.



When x and y values of the data pairs are exchanged, the sample correlation coefficient r remains the same. Problem 20 explores this result.

Interpretation It can be shown mathematically that r is always a number between $+1$ and -1 ($-1 \leq r \leq +1$). Table 9-2 gives a quick summary of some basic facts about r .

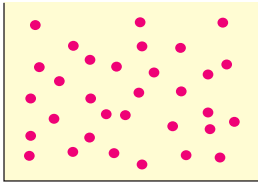
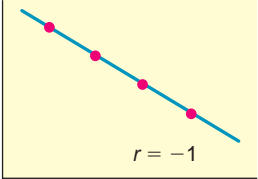
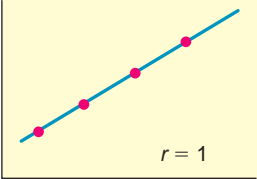
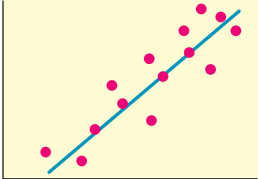
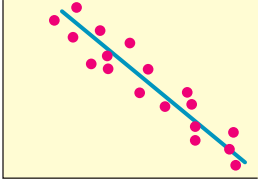
For most applications, you will use a calculator or computer software to compute r directly. However, to build some familiarity with the structure of the sample correlation coefficient, it is useful to do some calculations for yourself. Example 2 and Guided Exercise 3 show how to use the computation formula to compute r .

EXAMPLE 2 COMPUTING r



Sand driven by wind creates large, beautiful dunes at the Great Sand Dunes National Monument, Colorado. Of course, the same natural forces also create large dunes in the Great Sahara and Arabia. Is there a linear correlation between wind velocity and sand drift rate? Let x be a random variable representing wind velocity (in 10 cm/sec) and let y be a random variable representing drift rate of sand (in 100 gm/cm/sec). A test site at the Great Sand Dunes National Monument gave the following information about x and y (Reference: *Hydrologic, Geologic, and Biologic Research at Great Sand Dunes National Monument*, Proceedings of the National Park Service Research Symposium).

TABLE 9-2 Some Facts about the Correlation Coefficient

If r Is	Then	The Scatter Diagram Might Look Something Like	
0	There is no linear relation among the points of the scatter diagram.		
1 or -1	There is a perfect linear relation between x and y values; all points lie on the least-squares line.		
Between 0 and 1 ($0 < r < 1$)	The x and y values have a <i>positive correlation</i> . By this, we mean that <i>large</i> x values are associated with <i>large</i> y values, and <i>small</i> x values are associated with <i>small</i> y values.		As we go from left to right, the least-squares line goes <i>up</i> .
Between -1 and 0 ($-1 < r < 0$)	The x and y values have a <i>negative correlation</i> . By this, we mean that <i>large</i> x values are associated with <i>small</i> y values, and <i>small</i> x values are associated with <i>large</i> y values.		As we go from left to right, the least-squares line goes <i>down</i> .

x	70	115	105	82	93	125	88
y	3	45	21	7	16	62	12

(a) Construct a scatter diagram. Do you expect r to be positive?

SOLUTION: Figure 9-7 displays the scatter diagram. From the scatter diagram, it appears that as x values increase, y values also tend to increase. Therefore, r should be positive.

FIGURE 9-7

Wind Velocity (10 cm/sec) and Drift Rate of Sand (100 gm/cm/sec)

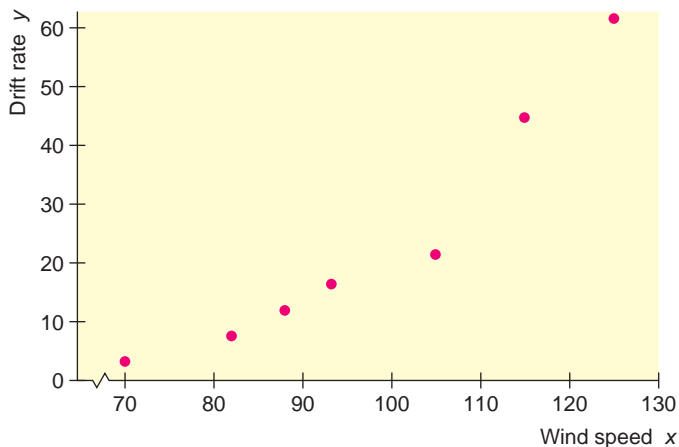


TABLE 9-3 Computation Table

x	y	x^2	y^2	xy
70	3	4900	9	210
115	45	13,225	2025	5175
105	21	11,025	441	2205
82	7	6724	49	574
93	16	8649	256	1488
125	62	15,625	3844	7750
88	12	7744	144	1056
$\Sigma x = 678$	$\Sigma y = 166$	$\Sigma x^2 = 67,892$	$\Sigma y^2 = 6768$	$\Sigma xy = 18,458$

(b) Compute r using the computation formula (formula 2).

SOLUTION: To find r , we need to compute Σx , Σx^2 , Σy , Σy^2 , and Σxy . It is convenient to organize the data in a table of five columns (Table 9-3) and then sum the entries in each column. Of course, many calculators give these sums directly. Using the computation formula for r , the sums from Table 9-3, and $n = 7$, we have

$$r = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{n\Sigma x^2 - (\Sigma x)^2} \sqrt{n\Sigma y^2 - (\Sigma y)^2}} \quad (2)$$

$$= \frac{7(18,458) - (678)(166)}{\sqrt{7(67,892) - (678)^2} \sqrt{7(6768) - (166)^2}} \approx \frac{16,658}{(124.74)(140.78)} \approx 0.949$$

Note: Using a calculator to compute r directly gives 0.949 to three places after the decimal.

(c) **Interpretation** What does the value of r tell you?

SOLUTION: Since r is very close to 1, we have an indication of a strong positive linear correlation between wind velocity and drift rate of sand. In other words, we expect that higher wind speeds tend to mean greater drift rates. Because r is so close to 1, the association between the variables appears to be linear.

Because it is quite a task to compute r for even seven data pairs, the use of columns as in Example 2 is extremely helpful. Your value for r should always be between -1 and 1 , inclusive. Use a scatter diagram to get a rough idea of the value of r . If your computed value of r is outside the allowable range, or if it disagrees quite a bit with the scatter diagram, recheck your calculations. Be sure you distinguish between expressions such as (Σx^2) and $(\Sigma x)^2$. Negligible rounding errors may occur, depending on how you (or your calculator) round.

GUIDED EXERCISE 3

Computing r

In one of the Boston city parks, there has been a problem with muggings in the summer months. A police cadet took a random sample of 10 days (out of the 90-day summer) and compiled the following data. For each day, x represents the number of police officers on duty in the park and y represents the number of reported muggings on that day.

x	10	15	16	1	4	6	18	12	14	7
y	5	2	1	9	7	8	1	5	3	6

Continued

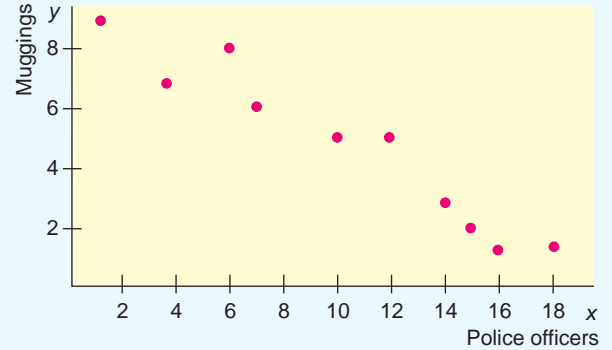
GUIDED EXERCISE 3 *continued*

(a) Construct a scatter diagram of x and y values.



Figure 9-8 shows the scatter diagram.

FIGURE 9-8 Scatter Diagram for Number of Police Officers versus Number of Muggings



(b) From the scatter diagram, do you think the computed value of r will be positive, negative, or zero? Explain.



r will be negative. The general trend is that large x values are associated with small y values, and vice versa. From left to right, the least-squares line goes down.

(c) Complete TABLE 9-4.



x	y	x^2	y^2	xy	
10	5	100	25	50	
15	2	225	4	30	
16	1	256	1	16	
1	9	1	81	9	
4	7	16	49	28	
6	8	_____	_____	_____	
18	1	_____	_____	_____	
12	5	_____	_____	_____	
14	3	_____	_____	_____	
7	6	49	36	42	
$\Sigma x = 103$		$\Sigma y = 47$	$\Sigma x^2 = \underline{\hspace{1cm}}$	$\Sigma y^2 = \underline{\hspace{1cm}}$	$\Sigma xy = \underline{\hspace{1cm}}$
$(\Sigma x)^2 = \underline{\hspace{1cm}}$		$(\Sigma y)^2 = \underline{\hspace{1cm}}$			

TABLE 9-5 Completion of Table 9-4

x	y	x^2	y^2	xy
6	8	36	64	48
18	1	324	1	18
12	5	144	25	60
14	3	196	9	42
		$\Sigma x^2 = 1347$	$\Sigma y^2 = 295$	$\Sigma xy = 343$
$(\Sigma x)^2 = 10,609$		$(\Sigma y)^2 = 2209$		

(d) Compute r . Alternatively, find the value of r directly by using a calculator or computer software.



$$\begin{aligned}
 r &= \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{n\Sigma x^2 - (\Sigma x)^2} \sqrt{n\Sigma y^2 - (\Sigma y)^2}} \\
 &= \frac{10(343) - (103)(47)}{\sqrt{10(1347) - (103)^2} \sqrt{10(295) - (47)^2}} \\
 &\approx \frac{-1411}{(53.49)(27.22)} \approx -0.969
 \end{aligned}$$

(e) **Interpretation** What does the value of r tell you about the relationship between the number of police officers and the number of muggings in the park?

There is a strong negative linear relationship between the number of police officers and the number of muggings. It seems that the more officers there are in the park, the fewer the number of muggings.

LOOKING FORWARD

If the scatter diagram and the value of the sample correlation coefficient r indicate a linear relationship between the data pairs, how do we find a suitable linear equation for the data? This process, called linear regression, is presented in the next section, Section 9.2.

TECH NOTES

Most calculators that support two-variable statistics provide the value of the sample correlation coefficient r directly. Statistical software provides r , r^2 , or both.

TI-84Plus/TI-83Plus/TI-nspire (with TI-84Plus keypad) First use CATALOG, find DiagnosticOn, and press Enter twice. Then, when you use STAT, CALC, option 8:LinReg(a + bx), the value of r will be given (data from Example 2). In the next section, we will discuss the line $y = a + bx$ and the meaning of r^2 .

Excel 2007 Excel gives the value of the sample correlation coefficient r in several outputs. One way to find the value of r is to click the Insert function (fx). Then in the dialogue box, select Statistical for the category and Correl for the function.

Minitab Use the menu selection Stat ► Basic Statistics ► Correlation.

```
LinReg
y=a+bx
a=-79.97763496
b=1.070565553
r2=.8997719968
r=.9485631222
```

LOOKING FORWARD

When the data are ranks (without ties) instead of measurements, the Pearson product-moment correlation coefficient can be reduced to a simpler equation called the Spearman rank correlation coefficient. This coefficient is used with nonparametric methods and is discussed in Section 11.3.

CRITICAL THINKING

Sample correlation compared to population correlation

Cautions about Correlation

The correlation coefficient can be thought of as a measure of how well a linear model fits the data points on a scatter diagram. The closer r is to $+1$ or -1 , the better a line “fits” the data. Values of r close to 0 indicate a poor fit to any line.

Usually a scatter diagram does not contain *all* possible data points that could be gathered. Most scatter diagrams represent only a *random sample* of data pairs taken from a very large population of all possible pairs. Because r is computed on the basis of a random sample of (x, y) pairs, we expect the values of r to vary from one sample to the next (much as the sample mean \bar{x} varies from sample to sample). This brings up the question of the *significance* of r . Or, put another way, what are the chances that our random sample of data pairs indicates a high correlation when, in fact, the population’s x and y values are not so strongly correlated? Right now, let’s just say that the significance of r is a separate issue that will be treated in Section 9.3, where we test the *population correlation coefficient* ρ (Greek letter *rho*, pronounced “row”).



Problem 21 demonstrates an informal process for determining whether or not r is significant. Problem 22 explores the effect of sample size on the significance of r . Problem 24 uses the value of ρ between two dependent variables to find the mean and standard deviation of a linear combination of the two variables.

Extrapolation

Causation

Lurking variables

r = **sample** correlation coefficient computed from a random sample of (x, y) data pairs.

ρ = **population** correlation coefficient computed from all population data pairs (x, y) .

There is a less formal way to address the significance of r using a table of “critical values” or “cut-off values” based on the r distribution and the number of data pairs. Problem 21 at the end of this section discusses this method.

The value of the sample correlation coefficient r and the strength of the linear relationship between variables is computed based on the sample data. The situation may change for measurements larger than or smaller than the data values included in the sample. For instance, for infants, there may be a high positive correlation between age in months and weight. However, that correlation might not apply for people ages 20 to 30 years.

The correlation coefficient is a mathematical tool for measuring the strength of a linear relationship between two variables. As such, it makes no implication about cause or effect. The fact that two variables tend to increase or decrease together does not mean that a change in one is *causing* a change in the other. A strong correlation between x and y is sometimes due to other (either known or unknown) variables. Such variables are called *lurking variables*.

In ordered pairs (x, y) , x is called the **explanatory** variable and y is called the **response** variable. When r indicates a linear correlation between x and y , changes in values of y tend to respond to changes in values of x according to a linear model. A **lurking variable** is a variable that is neither an explanatory nor a response variable. Yet, a lurking variable may be responsible for changes in both x and y .

EXAMPLE 3

CAUSATION AND LURKING VARIABLES

Over a period of years, the population of a certain town increased. It was observed that during this period the correlation between x , the number of people attending church, and y , the number of people in the city jail, was $r = 0.90$. Does going to church *cause* people to go to jail? Is there a *lurking variable* that might cause both variables x and y to increase?

SOLUTION: We hope church attendance does not cause people to go to jail! During this period, there was an increase in population. Therefore, it is not too surprising that both the number of people attending church and the number of people in jail increased. The high correlation between x and y is likely due to the lurking variable of population increase.

Correlation between averages



Problem 23 at the end of this section explores the correlation of averages.

The correlation between two variables consisting of averages is usually higher than the correlation between two variables representing corresponding raw data. One reason is that the use of averages reduces the variation that exists between individual measurements (see Section 6.5 and the central limit theorem). A high correlation based on two variables consisting of averages does not necessarily imply a high correlation between two variables consisting of individual measurements. See Problem 23 at the end of this section.

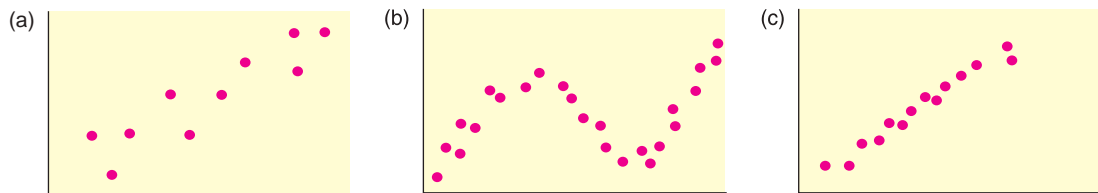
VIEWPOINT**Low on Credit, High on Cost!!!**

How do you measure automobile insurance risk? One way is to use a little statistics and customer credit ratings. Insurers say statistics show that drivers who have a history of bad credit are more likely to be in serious car accidents. According to a high-level executive at Allstate Insurance Company, financial instability is an extremely powerful predictor of future insurance losses. In short, there seems to be a strong correlation between bad credit ratings and auto insurance claims. Consequently, insurance companies want to charge higher premiums to customers with bad credit ratings. Consumer advocates object strongly because they say bad credit does not cause automobile accidents, and more than 20 states prohibit or restrict the use of credit ratings to determine auto insurance premiums. Insurance companies respond by saying that your best defense is to pay your bills on time!

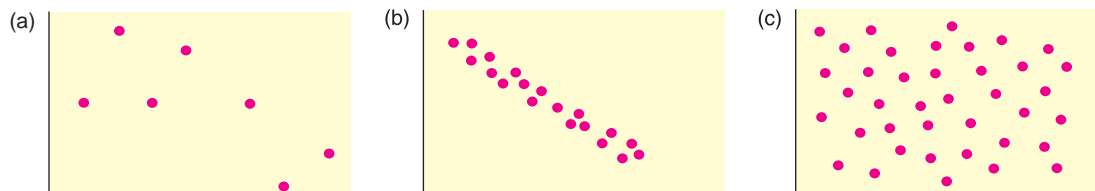
**SECTION 9.1
PROBLEMS**

Note: Answers may vary due to rounding.

- Statistical Literacy** When drawing a scatter diagram, along which axis is the explanatory variable placed? Along which axis is the response variable placed?
- Statistical Literacy** Suppose two variables are positively correlated. Does the response variable increase or decrease as the explanatory variable increases?
- Statistical Literacy** Suppose two variables are negatively correlated. Does the response variable increase or decrease as the explanatory variable increases?
- Statistical Literacy** Describe the relationship between two variables when the correlation coefficient r is
 - near -1 .
 - near 0 .
 - near 1 .
- Critical Thinking: Linear Correlation** Look at the following diagrams. Does each diagram show high linear correlation, moderate or low linear correlation, or no linear correlation?



- Critical Thinking: Linear Correlation** Look at the following diagrams. Does each diagram show high linear correlation, moderate or low linear correlation, or no linear correlation?



- Critical Thinking: Lurking Variables** Over the past few years, there has been a strong positive correlation between the annual consumption of diet soda drinks and the number of traffic accidents.

- (a) Do you think increasing consumption of diet soda drinks causes traffic accidents? Explain.
- (b) What lurking variables might be causing the increase in one or both of the variables? Explain.
8. **Critical Thinking: Lurking Variables** Over the past decade, there has been a strong positive correlation between teacher salaries and prescription drug costs.
- (a) Do you think paying teachers more causes prescription drugs to cost more? Explain.
- (b) What lurking variables might be causing the increase in one or both of the variables? Explain.
9. **Critical Thinking: Lurking Variables** Over the past 50 years, there has been a strong negative correlation between average annual income and the record time to run 1 mile. In other words, average annual incomes have been rising while the record time to run 1 mile has been decreasing.
- (a) Do you think increasing incomes cause decreasing times to run the mile? Explain.
- (b) What lurking variables might be causing the increase in one or both of the variables? Explain.
10. **Critical Thinking: Lurking Variables** Over the past 30 years in the United States, there has been a strong negative correlation between the number of infant deaths at birth and the number of people over age 65.
- (a) Is the fact that people are living longer causing a decrease in infant mortalities at birth?
- (b) What lurking variables might be causing the increase in one or both of the variables? Explain.
11. **Interpretation** Trevor conducted a study and found that the correlation between the price of a gallon of gasoline and gasoline consumption has a linear correlation coefficient of -0.7 . What does this result say about the relationship between price of gasoline and consumption? The study included gasoline prices ranging from \$2.70 to \$5.30 per gallon. Is it reliable to apply the results of this study to prices of gasoline higher than \$5.30 per gallon? Explain.
12. **Interpretation** Do people who spend more time on social networking sites spend more time using Twitter? Megan conducted a study and found that the correlation between the times spent on the two activities was 0.8. What does this result say about the relationship between times spent on the two activities? If someone spends more time than average on a social networking site, can you automatically conclude that he or she spends more time than average using Twitter? Explain.
13. **Veterinary Science: Shetland Ponies** How much should a healthy Shetland pony weigh? Let x be the age of the pony (in months), and let y be the average weight of the pony (in kilograms). The following information is based on data taken from *The Merck Veterinary Manual* (a reference used in most veterinary colleges).

x	3	6	12	18	24
y	60	95	140	170	185

- (a) Make a scatter diagram and draw the line you think best fits the data.
- (b) Would you say the correlation is low, moderate, or strong? positive or negative?
- (c) Use a calculator to verify that $\Sigma x = 63$, $\Sigma x^2 = 1089$, $\Sigma y = 650$, $\Sigma y^2 = 95,350$, and $\Sigma xy = 9930$. Compute r . As x increases from 3 to 24 months, does the value of r imply that y should tend to increase or decrease? Explain.

14. **Health Insurance: Administrative Cost** The following data are based on information from *Domestic Affairs*. Let x be the average number of employees in a group health insurance plan, and let y be the average administrative cost as a percentage of claims.

x	3	7	15	35	75
y	40	35	30	25	18

- (a) Make a scatter diagram and draw the line you think best fits the data.
 (b) Would you say the correlation is low, moderate, or strong? positive or negative?
 (c) Use a calculator to verify that $\Sigma x = 135$, $\Sigma x^2 = 7133$, $\Sigma y = 148$, $\Sigma y^2 = 4674$, and $\Sigma xy = 3040$. Compute r . As x increases from 3 to 75, does the value of r imply that y should tend to increase or decrease? Explain.
15. **Meteorology: Cyclones** Can a low barometer reading be used to predict maximum wind speed of an approaching tropical cyclone? Data for this problem are based on information taken from *Weatherwise* (Vol. 46, No. 1), a publication of the American Meteorological Society. For a random sample of tropical cyclones, let x be the lowest pressure (in millibars) as a cyclone approaches, and let y be the maximum wind speed (in miles per hour) of the cyclone.

x	1004	975	992	935	985	932
y	40	100	65	145	80	150

- (a) Make a scatter diagram and draw the line you think best fits the data.
 (b) Would you say the correlation is low, moderate, or strong? positive or negative?
 (c) Use a calculator to verify that $\Sigma x = 5823$, $\Sigma x^2 = 5,655,779$, $\Sigma y = 580$, $\Sigma y^2 = 65,750$, and $\Sigma xy = 556,315$. Compute r . As x increases, does the value of r imply that y should tend to increase or decrease? Explain.
16. **Geology: Earthquakes** Is the magnitude of an earthquake related to the depth below the surface at which the quake occurs? Let x be the magnitude of an earthquake (on the Richter scale), and let y be the depth (in kilometers) of the quake below the surface at the epicenter. The following is based on information taken from the National Earthquake Information Service of the U.S. Geological Survey. Additional data may be found by visiting the Brase/Brase statistics site at <http://www.cengage.com/statistics/brase> and finding the link to Earthquakes.

x	2.9	4.2	3.3	4.5	2.6	3.2	3.4
y	5.0	10.0	11.2	10.0	7.9	3.9	5.5

- (a) Make a scatter diagram and draw the line you think best fits the data.
 (b) Would you say the correlation is low, moderate, or strong? positive or negative?
 (c) Use a calculator to verify that $\Sigma x = 24.1$, $\Sigma x^2 = 85.75$, $\Sigma y = 53.5$, $\Sigma y^2 = 458.31$, and $\Sigma xy = 190.18$. Compute r . As x increases, does the value of r imply that y should tend to increase or decrease? Explain.
17. **Baseball: Batting Averages and Home Runs** In baseball, is there a linear correlation between batting average and home run percentage? Let x represent the batting average of a professional baseball player, and let y represent the player's home run percentage (number of home runs per 100 times at bat). A random sample of $n = 7$ professional baseball players gave the following information (Reference: *The Baseball Encyclopedia*, Macmillan Publishing Company).

x	0.243	0.259	0.286	0.263	0.268	0.339	0.299
y	1.4	3.6	5.5	3.8	3.5	7.3	5.0

- (a) Make a scatter diagram and draw the line you think best fits the data.
- (b) Would you say the correlation is low, moderate, or high? positive or negative?
- (c) Use a calculator to verify that $\Sigma x = 1.957$, $\Sigma x^2 \approx 0.553$, $\Sigma y = 30.1$, $\Sigma y^2 = 150.15$, and $\Sigma xy \approx 8.753$. Compute r . As x increases, does the value of r imply that y should tend to increase or decrease? Explain.
18. **University Crime: FBI Report** Do larger universities tend to have more property crime? University crime statistics are affected by a variety of factors. The surrounding community, accessibility given to outside visitors, and many other factors influence crime rates. Let x be a variable that represents student enrollment (in thousands) on a university campus, and let y be a variable that represents the number of burglaries in a year on the university campus. A random sample of $n = 8$ universities in California gave the following information about enrollments and annual burglary incidents (Reference: *Crime in the United States*, Federal Bureau of Investigation).

x	12.5	30.0	24.5	14.3	7.5	27.7	16.2	20.1
y	26	73	39	23	15	30	15	25



- (a) Make a scatter diagram and draw the line you think best fits the data.
- (b) Would you say the correlation is low, moderate, or high? positive or negative?
- (c) Using a calculator, verify that $\Sigma x = 152.8$, $\Sigma x^2 = 3350.98$, $\Sigma y = 246$, $\Sigma y^2 = 10,030$, and $\Sigma xy = 5488.4$. Compute r . As x increases, does the value of r imply that y should tend to increase or decrease? Explain.
19. **Expand Your Knowledge: Effect of Scale on Scatter Diagram** The initial visual impact of a scatter diagram depends on the scales used on the x and y axes. Consider the following data:

x	1	2	3	4	5	6
y	1	4	6	3	6	7



- (a) Make a scatter diagram using the same scale on both the x and y axes (i.e., make sure the unit lengths on the two axes are equal).
- (b) Make a scatter diagram using a scale on the y axis that is twice as long as that on the x axis.
- (c) Make a scatter diagram using a scale on the y axis that is half as long as that on the x axis.
- (d) On each of the three graphs, draw the straight line that you think best fits the data points. How do the slopes (or directions) of the three lines appear to change? *Note:* The actual slopes will be the same; they just appear different because of the choice of scale factors.
20. **Expand Your Knowledge: Effect on r of Exchanging x and y Values** Examine the computation formula for r , the sample correlation coefficient [formulas (1) and (2) of this section].
- (a) In the formula for r , if we exchange the symbols x and y , do we get a different result or do we get the same (equivalent) result? Explain.
- (b) If we have a set of x and y data values and we exchange corresponding x and y values to get a new data set, should the sample correlation coefficient be the same for both sets of data? Explain.

- (c) Compute the sample correlation coefficient r for each of the following data sets and show that r is the same for both.

x	1	3	4
y	2	1	6

x	2	1	6
y	1	3	4



21. **Expand Your Knowledge: Using a Table to Test ρ** The correlation coefficient r is a *sample* statistic. What does it tell us about the value of the population correlation coefficient ρ (Greek letter rho)? We will build the formal structure of hypothesis tests of ρ in Section 9.3. However, there is a quick way to determine if the sample evidence based on r is strong enough to conclude that there is some population correlation between the variables. In other words, we can use the value of r to determine if $\rho \neq 0$. We do this by comparing the value $|r|$ to an entry in Table 9-6. The value of α in the table gives us the probability of concluding that $\rho \neq 0$ when, in fact, $\rho = 0$ and there is no population correlation. We have two choices for α : $\alpha = 0.05$ or $\alpha = 0.01$.

PROCEDURE

HOW TO USE TABLE 9-6 TO TEST ρ

1. First compute r from a random sample of n data pairs (x, y) .
2. Find the table entry in the row headed by n and the column headed by your choice of α . Your choice of α is the risk you are willing to take of mistakenly concluding that $\rho \neq 0$ when, in fact, $\rho = 0$.
3. Compare $|r|$ to the table entry.
 - (a) If $|r| \geq$ table entry, then there is sufficient evidence to conclude that $\rho \neq 0$, and we say that r is **significant**. In other words, we conclude that there is some population correlation between the two variables x and y .
 - (b) If $|r| <$ table entry, then the evidence is insufficient to conclude that $\rho \neq 0$, and we say that r is **not significant**. We do not have enough evidence to conclude that there is any correlation between the two variables x and y .

TABLE 9-6 Critical Values for Correlation Coefficient r

n	$\alpha = 0.05$	$\alpha = 0.01$	n	$\alpha = 0.05$	$\alpha = 0.01$	n	$\alpha = 0.05$	$\alpha = 0.01$
3	1.00	1.00	13	0.53	0.68	23	0.41	0.53
4	0.95	0.99	14	0.53	0.66	24	0.40	0.52
5	0.88	0.96	15	0.51	0.64	25	0.40	0.51
6	0.81	0.92	16	0.50	0.61	26	0.39	0.50
7	0.75	0.87	17	0.48	0.61	27	0.38	0.49
8	0.71	0.83	18	0.47	0.59	28	0.37	0.48
9	0.67	0.80	19	0.46	0.58	29	0.37	0.47
10	0.63	0.76	20	0.44	0.56	30	0.36	0.46
11	0.60	0.73	21	0.43	0.55			
12	0.58	0.71	22	0.42	0.54			

- (a) Look at Problem 13 regarding the variables x = age of a Shetland pony and y = weight of that pony. Is the value of $|r|$ large enough to conclude that weight and age of Shetland ponies are correlated? Use $\alpha = 0.05$.

- (b) Look at Problem 15 regarding the variables x = lowest barometric pressure as a cyclone approaches and y = maximum wind speed of the cyclone. Is the value of $|r|$ large enough to conclude that lowest barometric pressure and wind speed of a cyclone are correlated? Use $\alpha = 0.01$.



22. **Expand Your Knowledge: Sample Size and Significance of Correlation** In this problem, we use Table 9-6 to explore the significance of r based on different sample sizes. See Problem 21.

- (a) Is a sample correlation coefficient $r = 0.820$ significant at the $\alpha = 0.01$ level based on a sample size of $n = 7$ data pairs? What about $n = 9$ data pairs?
 (b) Is a sample correlation coefficient $r = 0.40$ significant at the $\alpha = 0.05$ level based on a sample size of $n = 20$ data pairs? What about $n = 27$ data pairs?
 (c) Is it true that in order to be significant, an r value must be larger than 0.90? larger than 0.70? larger than 0.50? What does sample size have to do with the significance of r ? Explain.



23. **Expand Your Knowledge: Correlation of Averages** Fuming because you are stuck in traffic? Roadway congestion is a costly item, in both time wasted and fuel wasted. Let x represent the *average* annual hours per person spent in traffic delays and let y represent the *average* annual gallons of fuel wasted per person in traffic delays. A random sample of eight cities showed the following data (Reference: *Statistical Abstract of the United States*, 122nd Edition).

x (hr)	28	5	20	35	20	23	18	5
y (gal)	48	3	34	55	34	38	28	9

- (a) Draw a scatter diagram for the data. Verify that $\Sigma x = 154$, $\Sigma x^2 = 3712$, $\Sigma y = 249$, $\Sigma y^2 = 9959$, and $\Sigma xy = 6067$. Compute r .

The data in part (a) represent *average* annual hours lost per person and *average* annual gallons of fuel wasted per person in traffic delays. Suppose that instead of using average data for different cities, you selected one person at random from each city and measured the annual number of hours lost x for that person and the annual gallons of fuel wasted y for the same person.

x (hr)	20	4	18	42	15	25	2	35
y (gal)	60	8	12	50	21	30	4	70

- (b) Compute \bar{x} and \bar{y} for both sets of data pairs and compare the averages. Compute the sample standard deviations s_x and s_y for both sets of data pairs and compare the standard deviations. In which set are the standard deviations for x and y larger? Look at the defining formula for r , Equation 1. Why do smaller standard deviations s_x and s_y tend to increase the value of r ?
 (c) Make a scatter diagram for the second set of data pairs. Verify that $\Sigma x = 161$, $\Sigma x^2 = 4583$, $\Sigma y = 255$, $\Sigma y^2 = 12,565$, and $\Sigma xy = 7071$. Compute r .
 (d) Compare r from part (a) with r from part (c). Do the data for averages have a higher correlation coefficient than the data for individual measurements? List some reasons why you think hours lost per individual and fuel wasted per individual might vary more than the same quantities averaged over all the people in a city.



24. **Expand Your Knowledge: Dependent Variables** In Section 5.1, we studied linear combinations of *independent* random variables. What happens if the variables are not independent? A lot of mathematics can be used to prove the following:

Covariance

Let x and y be random variables with means μ_x and μ_y , variances σ_x^2 and σ_y^2 , and population correlation coefficient ρ (the Greek letter rho). Let a and b be any constants and let $w = ax + by$. Then,

$$\begin{aligned}\mu_w &= a\mu_x + b\mu_y \\ \sigma_w^2 &= a^2\sigma_x^2 + b^2\sigma_y^2 + 2ab\sigma_x\sigma_y\rho\end{aligned}$$

In this formula, ρ is the population correlation coefficient, theoretically computed using the population of all (x, y) data pairs. The expression $\sigma_x\sigma_y\rho$ is called the *covariance* of x and y . If x and y are independent, then $\rho = 0$ and the formula for σ_w^2 reduces to the appropriate formula for independent variables (see Section 5.1). In most real-world applications, the population parameters are not known, so we use sample estimates with the understanding that our conclusions are also estimates.

Do you have to be rich to invest in bonds and real estate? No, mutual fund shares are available to you even if you aren't rich. Let x represent annual percentage return (after expenses) on the Vanguard Total Bond Index Fund, and let y represent annual percentage return on the Fidelity Real Estate Investment Fund. Over a long period of time, we have the following population estimates (based on *Morningstar Mutual Fund Report*).

$$\mu_x \approx 7.32 \quad \sigma_x \approx 6.59 \quad \mu_y \approx 13.19 \quad \sigma_y \approx 18.56 \quad \rho \approx 0.424$$

- Do you think the variables x and y are independent? Explain.
- Suppose you decide to put 60% of your investment in bonds and 40% in real estate. This means you will use a weighted average $w = 0.6x + 0.4y$. Estimate your expected percentage return μ_w and risk σ_w .
- Repeat part (b) if $w = 0.4x + 0.6y$.
- Compare your results in parts (b) and (c). Which investment has the higher expected return? Which has the greater risk as measured by σ_w ?

SECTION 9.2

Linear Regression and the Coefficient of Determination

FOCUS POINTS

- State the least-squares criterion.
- Use sample data to find the equation of the least-squares line. Graph the least-squares line.
- Use the least-squares line to predict a value of the response variable y for a specified value of the explanatory variable x .
- Explain the difference between interpolation and extrapolation.
- Explain why extrapolation beyond the sample data range might give results that are misleading or meaningless.
- Use r^2 to determine *explained* and *unexplained* variation of the response variable y .

In Denali National Park, Alaska, the wolf population is dependent on a large, strong caribou population. In this wild setting, caribou are found in very large herds. The well-being of an entire caribou herd is not threatened by wolves. In fact, it is thought that wolves keep caribou herds strong by helping prevent overpopulation. Can the caribou population be used to predict the size of the wolf population?

Let x be a random variable that represents the fall caribou population (in hundreds) in Denali National Park, and let y be a random variable that represents the late-winter wolf population in the park. A random sample of recent years gave

the following information (Reference: U.S. Department of the Interior, National Biological Service).

x	30	34	27	25	17	23	20
y	66	79	70	60	48	55	60

Looking at the scatter diagram in Figure 9-9, we can ask some questions.

1. Do the data indicate a linear relationship between x and y ?
2. Can you find an equation for the best-fitting line relating x and y ? Can you use this relationship to predict the size of the wolf population when you know the size of the caribou population?
3. What fractional part of the variability in y can be associated with the variability in x ? What fractional part of the variability in y is not associated with a corresponding variability in x ?

FIGURE 9-9

Caribou and Wolf Populations

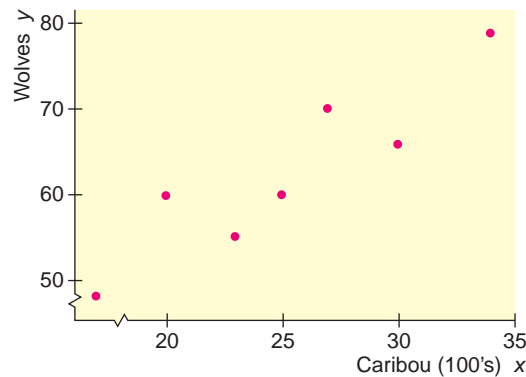
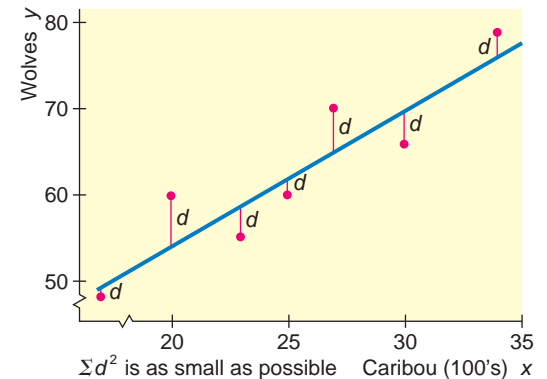


FIGURE 9-10

Least-Squares Criterion



The first step in answering these questions is to try to express the relationship as a mathematical equation. There are many possible equations, but the simplest and most widely used is the linear equation, or the equation of a straight line. Because we will be using this line to predict the y values from the x values, we call x the *explanatory variable* and y the *response variable*.

Our job is to find the linear equation that “best” represents the points of the scatter diagram. For our criterion of best-fitting line, we use the *least-squares criterion*, which states that the line we fit to the data points must be such that *the sum of the squares of the vertical distances from the points to the line be made as small as possible*. The least-squares criterion is illustrated in Figure 9-10.

Least-squares criterion

The sum of the squares of the vertical distances from the data points (x, y) to the line is made as small as possible.

In Figure 9-10, d represents the difference between the y coordinate of the data point and the corresponding y coordinate on the line. Thus, if the data point lies above the line, d is positive, but if the data point is below the line, d is negative. As a result, the sum of the d values can be small even if the points are widely spread in the scatter diagram. However, the squares d^2 cannot be negative. By minimizing the sum of the squares, we are, in effect, not allowing positive and negative d values to “cancel out” one another in the sum. It is in this way that we

Explanatory variable
Response variable
Least-squares criterion

can meet the least-squares criterion of minimizing the sum of the squares of the vertical distances between the points and the line over *all* points in the scatter diagram.

Least-squares line

We use the notation $\hat{y} = a + bx$ for the *least-squares line*. A little algebra tells us that b is the slope and a is the intercept of the line. In this context, \hat{y} (read “y hat”) represents the value of the response variable y estimated using the least-squares line and a given value of the explanatory variable x .

Techniques of calculus can be applied to show that a and b may be computed using the following procedure.

PROCEDURE

HOW TO FIND THE EQUATION OF THE LEAST-SQUARES LINE

$$\hat{y} = a + bx$$

Requirements for Statistical Inference

Obtain a random sample of n data pairs (x, y) , where x is the *explanatory variable* and y is the *response variable*. The data pairs should have a *bivariate normal distribution*. This means that for a fixed value of x , the y values should have a normal distribution (or at least a mound-shaped and symmetric distribution), and for a fixed y , the x values should have their own (approximately) normal distribution.

Procedure

1. Using the data pairs, compute Σx , Σy , Σx^2 , Σy^2 , and Σxy . Then compute the sample means \bar{x} and \bar{y} .
2. With $n =$ sample size, Σx , Σy , Σx^2 , Σy^2 , Σxy , \bar{x} , and \bar{y} , you are ready to compute the slope b and intercept a using the computation formulas

$$\text{Slope: } b = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{n\Sigma x^2 - (\Sigma x)^2} \tag{3}$$

$$\text{Intercept: } a = \bar{y} - b\bar{x} \tag{4}$$

Be careful! The notation Σx^2 means first square x and then calculate the sum, whereas $(\Sigma x)^2$ means first sum the x values and then square the result.

3. The equation of the least-squares line computed from your sample data is

$$\hat{y} = a + bx \tag{5}$$



Problem 21 demonstrates that for the same data set, the least-squares lines for predicting y or for predicting x are essentially different.

Slope b

Intercept a



For data following exponential growth or power law models, logarithmic transformations can be used to transform the data into linear models. Then linear regression can be used on the transformed data. Problems 22–25 show these methods.

COMMENT The computation formulas for the slope of the least-squares line, the sample correlation coefficient r , and the standard deviations s_x and s_y use many of the same sums. There is, in fact, a relationship between the sample correlation coefficient r and the slope of the least-squares line b . In instances where we know r , s_x , and s_y , we can use the following formula to compute b .

$$b = r \left(\frac{s_y}{s_x} \right) \tag{6}$$

COMMENT In other mathematics courses, the slope-intercept form of the equation of a line is usually given as $y = mx + b$, where m refers to the slope of the line and b to the y coordinate of the y intercept. In statistics, when there is only one explanatory variable, it is common practice to use the letter b to designate the slope of the least-squares line and the letter a to designate the y coordinate of the intercept. For example, these are the symbols used on the TI-84Plus/TI-83Plus/TI-*n*spire calculators as well as on many other calculators.

Using the formulas to find the values of a and b

For most applications, you can use a calculator or computer software to compute a and b directly. However, to build some familiarity with the structure of the computation formulas, it is useful to do some calculations yourself. Example 4 shows how to use the computation formulas to find the values of a and b and the equation of the least-squares line $\hat{y} = a + bx$.

Note: If you are using your calculator to find the values of a and b directly, you may omit the discussion regarding use of the formulas. Go to the margin header “Using the values of a and b to construct the equation of the least-squares line.”

EXAMPLE 4 LEAST-SQUARES LINE

Let's find the least-squares equation relating the variables x = size of caribou population (in hundreds) and y = size of wolf population in Denali National Park. Use x as the explanatory variable and y as the response variable.



Joe McDonald/Encyclopedia/Corbis

- (a) Use the computation formulas to find the slope b of the least-squares line and the y intercept a .

SOLUTION: Table 9-7 gives the data values x and y along with the values x^2 , y^2 , and xy . First compute the sample means.

$$\bar{x} = \frac{\Sigma x}{n} = \frac{176}{7} \approx 25.14 \quad \text{and} \quad \bar{y} = \frac{\Sigma y}{n} = \frac{438}{7} \approx 62.57$$

Next compute the slope b .

$$b = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{n\Sigma x^2 - (\Sigma x)^2} = \frac{7(11,337) - (176)(438)}{7(4628) - (176)^2} = \frac{2271}{1420} \approx 1.60$$

Use the values of b , \bar{x} and \bar{y} to compute the y intercept a .

$$a = \bar{y} - b\bar{x} \approx 62.57 - 1.60(25.14) \approx 22.35$$

Note that calculators give the values $b \approx 1.599$ and $a \approx 22.36$. These values differ slightly from those you computed using the formulas because of rounding.

- (b) Use the values of a and b (either computed or obtained from a calculator) to find the equation of the least-squares line.

SOLUTION:

$$\begin{aligned} \hat{y} &= a + bx \\ \hat{y} &\approx 22.35 + 1.60x \quad \text{since } a \approx 22.35 \quad \text{and} \quad b \approx 1.60 \end{aligned}$$

TABLE 9-7 Sums for Computing b , \bar{x} , and \bar{y}

x	y	x^2	y^2	xy
30	66	900	4356	1980
34	79	1156	6241	2686
27	70	729	4900	1890
25	60	625	3600	1500
17	48	289	2304	816
23	55	529	3025	1265
20	60	400	3600	1200
$\Sigma x = 176$	$\Sigma y = 438$	$\Sigma x^2 = 4628$	$\Sigma y^2 = 28,026$	$\Sigma xy = 11,337$

Graphing the least-squares line

(c) Graph the equation of the least-squares line on a scatter diagram.

SOLUTION: To graph the least-squares line, we have several options available. The slope-intercept method of algebra is probably the quickest, but may not always be convenient if the intercept is not within the range of the sample data values. It is just as easy to select two x values in the range of the x data values and then use the least-squares line to compute two corresponding \hat{y} values.

In fact, we already have the coordinates of one point on the least-squares line. By the formula for the intercept [Equation (4)], the point (\bar{x}, \bar{y}) is always on the least-squares line. For our example, $(\bar{x}, \bar{y}) = (25.14, 62.57)$.

The point (\bar{x}, \bar{y}) is always on the least-squares line.

Another x value within the data range is $x = 34$. Using the least-squares line to compute the corresponding \hat{y} value gives

$$\hat{y} \approx 22.35 + 1.60(34) \approx 76.75$$

We place the two points $(25.14, 62.57)$ and $(34, 76.75)$ on the scatter diagram (using a different symbol than that used for the sample data points) and connect the points with a line segment (Figure 9-11).

Meaning of slope

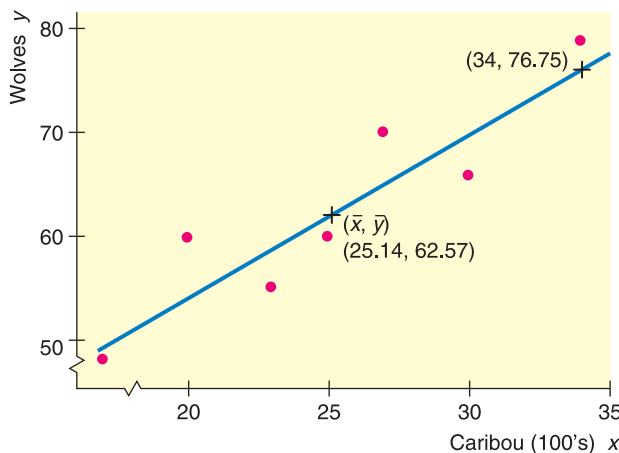
In the equation $\hat{y} = a + bx$, the slope b tells us how many units \hat{y} changes for each unit change in x . In Example 4 regarding size of wolf and caribou populations,

$$\hat{y} \approx 22.35 + 1.60x$$

The slope 1.60 tells us that if the number of caribou (in hundreds) changes by 1 (hundred), then we expect the sustainable wolf population to change by 1.60. In other words, our model says that an increase of 100 caribou will increase the predicted wolf population by 1.60. If the caribou population decreases by 400, we predict the sustainable wolf population to decrease by 6.4.

FIGURE 9-11

Caribou and Wolf Populations



The slope of the least-squares line tells us how many units the response variable is expected to change for each unit change in the explanatory variable. The number of units change in the response variable for each unit change in the explanatory variable is called the **marginal change** of the response variable.

Marginal change

Some points in the data set have a strong influence on the equation of the least-squares line.

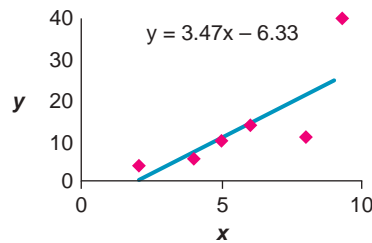
Influential points

A data pair is **influential** if removing it would substantially change the equation of the least-squares line or other calculations associated with linear regression. An influential point often has an x -value near the extreme high or low value of the data set.

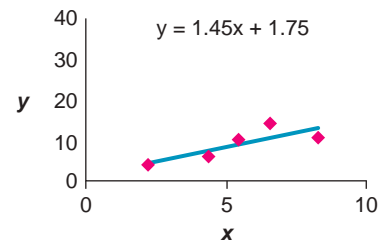
Figure 9-12 shows two scatter diagrams produced in Excel. Figure 9-12(a) has an influential point. Figure 9-12(b) shows the scatter diagram with the influential point removed. Notice that the equations for the least-squares lines are quite different.

FIGURE 9-12

(a) Influential Point Present



(b) Influential Point Removed



If a data set has an influential point, look at the influential point carefully to make sure it is not the result of a data collection error or a data-recording error. A valid influential point affects the equation of the least-squares line. The group project in Data Highlights at the end of this chapter further explores influential points.



CRITICAL THINKING

Predicting y for a specified x

Using the Least-Squares Line for Prediction

Making predictions is one of the main applications of linear regression. In other words, you use the equation of the least-squares line to predict the \hat{y} value for a specified x value. The accuracy of the prediction depends on several components.

How well does the least-squares line fit the original data points?

The accuracy of the prediction depends on how well the least-squares line fits the original raw data points. Here are some tools to assess the fit of the line.

- Look at the scatter diagram, taking into account the scale of each axis.
- See if there are any influential points.
- Consider the value of the sample correlation coefficient r . The closer r is to 1 or -1 , the better the least-squares line fits the data.
- Consider the value of the coefficient of determination r^2 (as discussed later in this section).
- Look at the residuals and a residual plot (see Problem 19 for a discussion of residual plots).

Residual



Problems 19 and 20 at the end of this section show how to make a residual plot.

The **residual** is the difference between the y value in a specified data pair (x, y) and the value $\hat{y} = a + bx$ predicted by the least-squares line for the same x .

$$y - \hat{y} \text{ is the residual}$$

If the residuals seem random about 0, the least-squares line provides a reasonable model for the data. Later in this section you will see the residual used in the

development of the *coefficient of determination*, another important measurement associated with linear regression.

Does the prediction involve interpolation or extrapolation?

Another issue that affects the validity of predictions is whether you are *interpolating* or *extrapolating*.

Predicting \hat{y} values for x values that are **between** observed x values in the data set is called **interpolation**.

Predicting \hat{y} values for x values that are **beyond** observed x values in the data set is called **extrapolation**. Extrapolation may produce unrealistic forecasts.

The least-squares line may not reflect the relationship between x and y for values of x outside the data range. For example, there is a fairly high correlation between height and age for boys ages 1 year to 10 years. In general, the older the boy, the taller the boy. A least-squares line based on such data would give good predictions of height for boys of ages between 1 and 10. However, it would be fairly meaningless to use the same linear regression line to predict the height of a 20-year-old or 50-year-old man.

The data are sample data.

Another consideration when working with predictions is the fact that the least-squares line is based on sample data. Each different sample will produce a slightly different equation for the least-squares line. Just as there are confidence intervals for parameters such as population means, there are confidence intervals for the prediction of y for a given x . We will examine confidence intervals for predictions in Section 9.3.

The least-squares line uses x as the explanatory variable and y as the response variables.

One more important fact about predictions: The least-squares line is developed with x as the explanatory variable and y as the response variable. This model can be used only to predict y values from specified x values. If you wish to begin with y values and predict corresponding x values, you must start all over and compute a new equation. Such an equation would be developed using a model with x as the response variable and y as the explanatory variable. See Problem 21 at the end of this section. Note that the equation for predicting x values *cannot* be derived from the least-squares line predicting y simply by solving the equation for x .

The least-squares line developed with x as the explanatory variable and y as the response variable can be used only to predict y values from specified x values.

The next example shows how to use the least-squares line for predictions.

EXAMPLE 5 PREDICTIONS

We continue with Example 4 regarding size of the wolf population as it relates to size of the caribou population. Suppose you want to predict the size of the wolf population when the size of the caribou population is 21 (hundred).



Joe McDonald/Corbis

- (a) In the least-squares model developed in Example 4, which is the explanatory variable and which is the response variable? Can you use the equation to predict the size of the wolf population for a specified size of caribou population?

SOLUTION: The least-squares line $\hat{y} \approx 22.35 + 1.60x$ was developed using $x =$ size of caribou population (in hundreds) as the explanatory variable and $y =$ size of wolf population as the response variable. We can use the equation to predict the y value for a specified x value.

- (b) The sample data pairs have x values ranging from 17 (hundred) to 34 (hundred) for the size of the caribou population. To predict the size of the wolf population when the size of the caribou population is 21 (hundred), will you be interpolating or extrapolating?

SOLUTION: Interpolating, since 21 (hundred) falls within the range of sample x values.

- (c) Predict the size of the wolf population when the caribou population is 21 (hundred).

SOLUTION: Using the least-squares line from Example 4 and the value 21 in place of x gives

$$\hat{y} \approx 22.35 + 1.60x \approx 22.35 + 1.60(21) \approx 55.95$$

Rounding up to a whole number gives a prediction of 56 for the size of the wolf population.

GUIDED EXERCISE 4

Least-squares line

The Quick Sell car dealership has been using 1-minute spot ads on a local TV station. The ads always occur during the evening hours and advertise the different models and price ranges of cars on the lot that week. During a 10-week period, a Quick Sell dealer kept a weekly record of the number x of TV ads versus the number y of cars sold. The results are given in Table 9-8.

The manager decided that Quick Sell can afford only 12 ads per week. At that level of advertisement, how many cars can Quick Sell expect to sell each week? We'll answer this question in several steps.

TABLE 9-8

x	y
6	15
20	31
0	10
14	16
25	28
16	20
28	40
18	25
10	12
8	15

- (a) Draw a scatter diagram for the data.

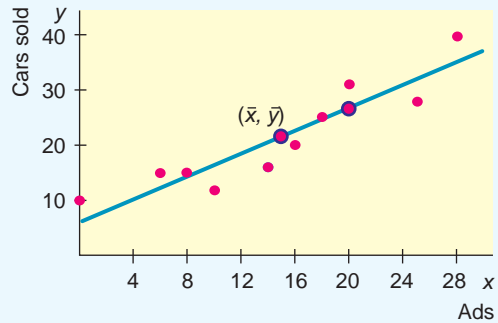


The scatter diagram is shown in Figure 9-13 on the next page. The plain red dots in Figure 9-13 are the points of the scatter diagram. Notice that the least-squares line is also shown with two extra points used to position that line.

Continued

GUIDED EXERCISE 4 *continued*

FIGURE 9-13 Scatter Diagram and Least-Squares Line for Table 9-8



(b) Look at Equations (3) to (5) pertaining to the least-squares line (page 522). Two of the quantities that we need to find b are (Σx) and (Σxy) . List the others.

➔ We also need n , (Σy) , (Σx^2) , and $(\Sigma x)^2$.

(c) Complete Table 9-9(a).

➔ The missing table entries are shown in Table 9-9(b).

TABLE 9-9(a)

x	y	x^2	xy
6	15	36	90
20	31	400	620
0	10	0	0
14	16	196	224
25	28	625	700
16	20	256	320
28	40	—	—
18	25	—	—
10	12	—	—
8	15	64	120
$\Sigma x = 145$	$\Sigma y = 212$	$\Sigma x^2 = \underline{\quad}$	$\Sigma xy = \underline{\quad}$

TABLE 9-9(b)

x^2	xy
$(28)^2 = 784$	$28(40) = 1120$
$(18)^2 = 324$	$18(25) = 450$
$(10)^2 = 100$	$10(12) = 120$
$\Sigma x^2 = 2785$	$\Sigma xy = 3764$

(d) Compute the sample means \bar{x} and \bar{y}

➔
$$\bar{x} = \frac{\Sigma x}{n} = \frac{145}{10} = 14.5$$

$$\bar{y} = \frac{\Sigma y}{n} = \frac{212}{10} = 21.2$$

(e) Compute a and b for the equation $\hat{y} = a + bx$ of the least-squares line.

➔
$$b = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{n\Sigma x^2 - (\Sigma x)^2}$$

$$= \frac{10(3764) - (145)(212)}{10(2785) - (145)^2} = \frac{6900}{6825} \approx 1.01$$

$$a = \bar{y} - b\bar{x}$$

$$\approx 21.2 - 1.01(14.5) \approx 6.56$$

(f) What is the equation of the least-squares line $\hat{y} = a + bx$?

➔ Using the values of a and b computed in part (e) or values of a and b obtained directly from a calculator,

$$\hat{y} \approx 6.56 + 1.01x$$

(g) Plot the least-squares line on your scatter diagram.

➔ The least-squares line goes through the point $(\bar{x}, \bar{y}) = (14.5, 21.2)$. To get another point on the line,

Continued

GUIDED EXERCISE 4 *continued*

- (h) Read the \hat{y} value for $x = 12$ from your graph. Then use the equation of the least-squares line to calculate \hat{y} when $x = 12$. How many cars can the manager expect to sell if 12 ads per week are aired on TV?

→ The graph gives $\hat{y} \approx 19$. From the equation, we get

$$\hat{y} \approx 6.56 + 1.01x$$

$$\approx 6.56 + 1.01(12) \quad \text{using 12 in place of } x$$

$$\approx 18.68$$

To the nearest whole number, the manager can expect to sell 19 cars when 12 ads are aired on TV each week.

- (i) **Interpretation** How reliable do you think the prediction is? Explain. (Guided Exercise 5 will show that $r \approx 0.919$.)

→ The prediction should be fairly reliable. The prediction involves interpolation, and the scatter diagram shows that the data points are clustered around the least-squares line. From the next Guided Exercise we have the information that r is high. Of course, other variables might affect the value of y for $x = 12$.


TECH NOTES

When we have more data pairs, it is convenient to use a technology tool such as the TI-84Plus/TI-83Plus/TI-*nspire* calculators, Excel 2007, or Minitab to find the equation of the least-squares line. The displays show results for the data of Guided Exercise 4 regarding car sales and ads.

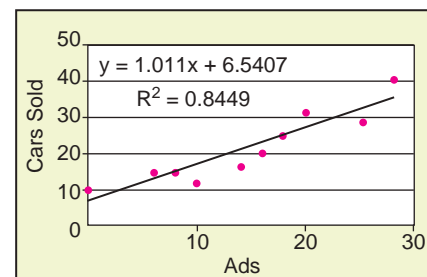
TI-84Plus/TI-83Plus/TI-*nspire* (with TI-84Plus keypad) Press **STAT**, choose **Calculate**, and use option **8:LinReg (a + bx)**. For a graph showing the scatter plot and the least-squares line, press the **STAT PLOT** key, turn on a plot, and highlight the first type. Then press the **Y =** key. To enter the equation of the least-squares line, press **VARS**, select **5:Statistics**, highlight **EQ**, and then highlight **1:RegEQ**. Press **ENTER**. Finally, press **ZOOM** and choose **9:ZoomStat**.

Excel 2007 There are several ways to find the equation of the least-squares line in Excel. One way is to make a scatter plot. On the home screen, click the **Insert** tab. In the Chart Group, select **Scatter** and choose the first type. In the next ribbon, the Chart Layout Group offers options for including titles and axes labels. Right click on any data point and select **Add Trendline**. In the dialogue box, select **linear** and check **Display Equation on Chart**. To display the value of the coefficient of determination, check **Display R-squared Value on Chart**.


TI-84Plus/TI-83/TI-*nspire* Display

```
LinReg
y=a+bx
a=6.5407
b=1.0110
r2=.8449
r=.9192
```

Excel 2007 Display



Minitab There are a number of ways to generate the least-squares line. One way is to use the menu selection **Stat** ► **Regression** ► **Fitted Line Plot**. The least-squares equation is shown with the diagram.



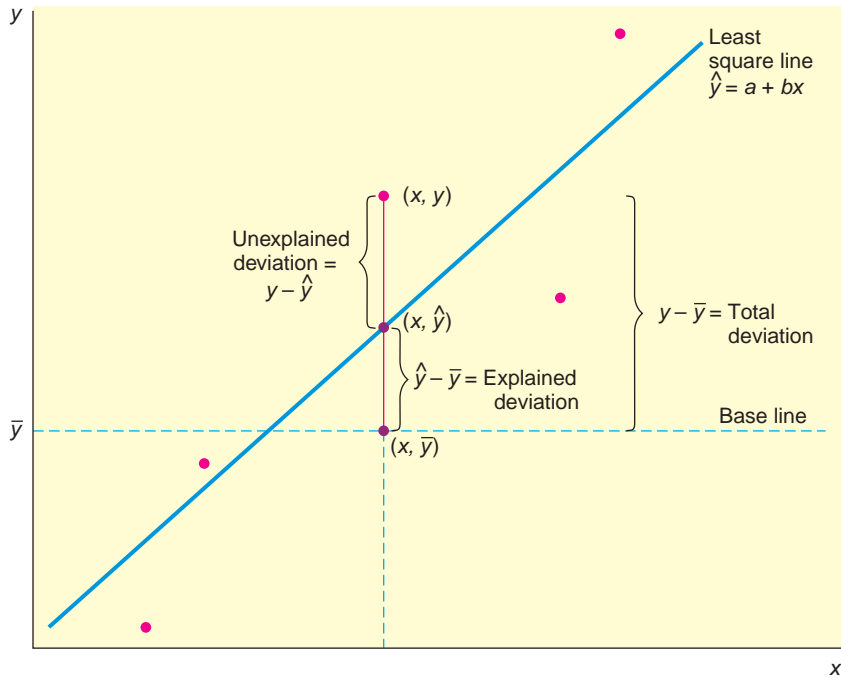
Coefficient of Determination

Coefficient of determination r^2

There is another way to answer the question “How good is the least-squares line as an instrument of regression?” The *coefficient of determination* r^2 is the square of the sample correlation coefficient r .

Suppose we have a scatter diagram and corresponding least-squares line, as shown in Figure 9-14.

FIGURE 9-14
Explained and Unexplained Deviations



Let us take the point of view that \bar{y} is a kind of baseline for the y values. If we were given an x value, and if we were completely ignorant of regression and correlation but we wanted to predict a value of y corresponding to the given x , a reasonable guess for y would be the mean \bar{y} . However, since we do know how to construct the least-squares regression line, we can calculate $\hat{y} = a + bx$, the predicted value corresponding to x . In most cases, the predicted value \hat{y} on the least-squares line will not be the same as the actual data value y . We will measure deviations (or differences) from the baseline \bar{y} . (See Figure 9-14.)

$$\begin{aligned} \text{Total deviation} &= y - \bar{y} \\ \text{Explained deviation} &= \hat{y} - \bar{y} \\ \text{Unexplained deviation} &= y - \hat{y} \quad (\text{also known as the } \textit{residual}) \end{aligned}$$

The total deviation $y - \bar{y}$ is a measure of how far y is from the baseline \bar{y} . This can be broken into two parts. The explained deviation $\hat{y} - \bar{y}$ tells us how far the estimated y value “should” be from the baseline \bar{y} . (The “explanation” of this part of the deviation is the least-squares line, so to speak.) The unexplained deviation $y - \hat{y}$ tells us how far our data value y is “off.” This amount is called *unexplained* because it is due to random chance and other factors that the least-squares line cannot account for.

$$\begin{pmatrix} y - \bar{y} \\ \text{Total} \\ \text{deviation} \end{pmatrix} = \begin{pmatrix} \hat{y} - \bar{y} \\ \text{Explained} \\ \text{deviation} \end{pmatrix} + \begin{pmatrix} y - \hat{y} \\ \text{Unexplained} \\ \text{deviation} \end{pmatrix}$$

At this point, we wish to include all the data pairs and we wish to deal only with nonnegative values (so that positive and negative deviations won’t cancel out).

Therefore, we construct the following equation for the sum of squares. This equation can be derived using some lengthy algebraic manipulations, which we omit.

Explained variation
Unexplained variation

$$\begin{aligned} \Sigma(y - \bar{y})^2 &= \Sigma(\hat{y} - \bar{y})^2 + \Sigma(y - \hat{y})^2 \\ \left(\begin{array}{c} \text{Total} \\ \text{variation} \end{array} \right) &= \left(\begin{array}{c} \text{Explained} \\ \text{variation} \end{array} \right) + \left(\begin{array}{c} \text{Unexplained} \\ \text{variation} \end{array} \right) \end{aligned}$$

Note that the sum of *squares* is taken over all data points and is then referred to as *variation* (not deviation).

The preceding concepts are connected together in the following important statement (whose proof we omit):

If r is the sample correlation coefficient [see Equation (2)], then it can be shown that

$$r^2 = \frac{\Sigma(\hat{y} - \bar{y})^2}{\Sigma(y - \bar{y})^2} = \frac{\text{Explained variation}}{\text{Total variation}}$$

r^2 is called the *coefficient of determination*.

Let us summarize our discussion.

Coefficient of determination r^2

1. Compute the sample correlation coefficient r using the procedure of Section 9.1. Then simply compute r^2 , the sample coefficient of determination.
2. **Interpretation** The value r^2 is the ratio of explained variation over total variation. That is, r^2 is the fractional amount of total variation in y that can be explained by using the linear model $\hat{y} = a + bx$.
3. **Interpretation** Furthermore, $1 - r^2$ is the fractional amount of total variation in y that is due to random chance or to the possibility of lurking variables that influence y .

In other words, the coefficient of determination r^2 is a measure of the proportion of variation in y that is explained by the regression line, using x as the explanatory variable. If $r = 0.90$, then $r^2 = 0.81$ is the coefficient of determination. We can say that about 81% of the (variation) behavior of the y variable can be explained by the corresponding (variation) behavior of the x variable if we use the equation of the least-squares line. The remaining 19% of the (variation) behavior of the y variable is due to random chance or to the possibility of lurking variables that influence y .

GUIDED EXERCISE 5

Coefficient of determination r^2

In Guided Exercise 4, we looked at the relationship between x = number of 1-minute spot ads on TV advertising different models of cars and y = number of cars sold each week by the sponsoring car dealership.

- (a) Using the sums found in Guided Exercise 4, compute the sample correlation coefficient r .
 $n = 10$, $\Sigma x = 145$, $\Sigma y = 212$, $\Sigma x^2 = 2785$, and $\Sigma xy = 3764$. You also need $\Sigma y^2 = 5320$.

$$r = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{n\Sigma x^2 - (\Sigma x)^2}\sqrt{n\Sigma y^2 - (\Sigma y)^2}}$$

$$\begin{aligned} \Rightarrow r &= \frac{10(3764) - (145)(212)}{\sqrt{10(2785) - (145)^2}\sqrt{10(5320) - (212)^2}} \\ &\approx \frac{6900}{(82.61)(90.86)} \\ &\approx 0.919 \end{aligned}$$

Continued

GUIDED EXERCISE 5 *continued*

- (b) Compute the coefficient of determination r^2 . $\Rightarrow r^2 \approx 0.845$
- (c) **Interpretation** What percentage of the variation in the number of car sales can be explained by the ads and the least-squares line? $\Rightarrow 84.5\%$
- (d) **Interpretation** What percentage of the variation in the number of car sales is not explained by the ads and the least-squares line? $\Rightarrow 100\% - 84.5\%$, or 15.5%

VIEWPOINT**It's Freezing!**

Can you use average temperatures in January to predict how bad the rest of the winter will be? Can you predict the number of days with freezing temperatures for the entire calendar year using conditions in January? How good would such a forecast be for predicting growing season or number of frost-free days? Methods of this section can help you answer such questions. For more information, visit the Brase/Brase statistics site at <http://www.cengage.com/statistics/brase> and find the link to *Temperatures*.

**SECTION 9.2
PROBLEMS**

- Statistical Literacy** In the least-squares line $\hat{y} = 5 - 2x$, what is the value of the slope? When x changes by 1 unit, by how much does \hat{y} change?
- Statistical Literacy** In the least squares line $\hat{y} = 5 + 3x$, what is the marginal change in \hat{y} for each unit change in x ?
- Critical Thinking** When we use a least-squares line to predict y values for x values beyond the range of x values found in the data, are we extrapolating or interpolating? Are there any concerns about such predictions?
- Critical Thinking** If two variables have a negative linear correlation, is the slope of the least-squares line positive or negative?
- Critical Thinking: Interpreting Computer Printouts** We use the form $\hat{y} = a + bx$ for the least-squares line. In some computer printouts, the least-squares equation is not given directly. Instead, the value of the constant a is given, and the coefficient b of the explanatory or predictor variable is displayed. Sometimes a is referred to as the constant, and sometimes as the intercept. Data from *Climatology Report No. 77-3* of the Department of Atmospheric Science, Colorado State University, showed the following relationship between elevation (in thousands of feet) and average number of frost-free days per year in Colorado locations.

A Minitab printout provides

Predictor	Coef	SE Coef	T	P
Constant	318.16	28.31	11.24	0.002
Elevation	-30.878	3.511	-8.79	0.003
s = 11.8603		R-Sq = 96.3%		

Notice that “Elevation” is listed under “Predictor.” This means that elevation is the explanatory variable x . Its coefficient is the slope b . “Constant” refers to a in the equation $\hat{y} = a + bx$.

- (a) Use the printout to write the least-squares equation.
- (b) For each 1000-foot increase in elevation, how many fewer frost-free days are predicted?
- (c) The printout gives the value of the coefficient of determination r^2 . What is the value of r ? Be sure to give the correct sign for r based on the sign of b .
- (d) **Interpretation** What percentage of the variation in y can be *explained* by the corresponding variation in x and the least-squares line? What percentage is *unexplained*?
6. **Critical Thinking: Interpreting Computer Printouts** Refer to the description of a computer display for regression described in Problem 5. The following Minitab display gives information regarding the relationship between the body weight of a child (in kilograms) and the metabolic rate of the child (in 100 kcal/24 hr). The data is based on information from *The Merck Manual* (a commonly used reference in medical schools and nursing programs).

Predictor	Coef	SE Coef	T	P
Constant	0.8565	0.4148	2.06	0.084
Weight	0.40248	0.02978	13.52	0.000

s = 0.517508 R-Sq = 96.8%

- (a) Write out the least-squares equation.
- (b) For each 1-kilogram increase in weight, how much does the metabolic rate of a child increase?
- (c) What is the value of the sample correlation coefficient r ?
- (d) **Interpretation** What percentage of the variation in y can be *explained* by the corresponding variation in x and the least-squares line? What percentage is *unexplained*?

For Problems 7–18, please do the following.

- (a) Draw a scatter diagram displaying the data.
- (b) Verify the given sums Σx , Σy , Σx^2 , Σy^2 , and Σxy and the value of the sample correlation coefficient r .
- (c) Find \bar{x} , \bar{y} , a , and b . Then find the equation of the least-squares line $\hat{y} = a + bx$.
- (d) Graph the least-squares line on your scatter diagram. Be sure to use the point (\bar{x}, \bar{y}) as one of the points on the line.
- (e) **Interpretation** Find the value of the coefficient of determination r^2 . What percentage of the variation in y can be *explained* by the corresponding variation in x and the least-squares line? What percentage is *unexplained*?
- Answers may vary slightly due to rounding.
7. **Economics: Entry-Level Jobs** An economist is studying the job market in Denver-area neighborhoods. Let x represent the total number of jobs in a given neighborhood, and let y represent the number of entry-level jobs in the same neighborhood. A sample of six Denver neighborhoods gave the following information (units in hundreds of jobs).

x	16	33	50	28	50	25
y	2	3	6	5	9	3

Source: *Neighborhood Facts*, The Piton Foundation. To find out more, visit the Brase/Brase statistics site at <http://www.cengage.com/statistics/brase> and find the link to the Piton Foundation.

Complete parts (a) through (e), given $\Sigma x = 202$, $\Sigma y = 28$, $\Sigma x^2 = 7754$, $\Sigma y^2 = 164$, $\Sigma xy = 1096$, and $r \approx 0.860$.

- (f) For a neighborhood with $x = 40$ jobs, how many are predicted to be entry-level jobs?

8. **Ranching: Cattle** You are the foreman of the Bar-S cattle ranch in Colorado. A neighboring ranch has calves for sale, and you are going to buy some to add to the Bar-S herd. How much should a healthy calf weigh? Let x be the age of the calf (in weeks), and let y be the weight of the calf (in kilograms). The following information is based on data taken from *The Merck Veterinary Manual* (a reference used by many ranchers).

x	1	3	10	16	26	26
y	42	50	75	100	150	200

Complete parts (a) through (e), given $\Sigma x = 92$, $\Sigma y = 617$, $\Sigma x^2 = 2338$, $\Sigma y^2 = 82,389$, $\Sigma xy = 13,642$, and $r \approx 0.998$.

- (f) The calves you want to buy are 12 weeks old. What does the least-squares line predict for a healthy weight?

9. **Weight of Car: Miles per Gallon** Do heavier cars really use more gasoline? Suppose a car is chosen at random. Let x be the weight of the car (in hundreds of pounds), and let y be the miles per gallon (mpg). The following information is based on data taken from *Consumer Reports* (Vol. 62, No. 4).

x	27	44	32	47	23	40	34	52
y	30	19	24	13	29	17	21	14

Complete parts (a) through (e), given $\Sigma x = 299$, $\Sigma y = 167$, $\Sigma x^2 = 11,887$, $\Sigma y^2 = 3773$, $\Sigma xy = 5814$, and $r \approx -0.946$.

- (f) Suppose a car weighs $x = 38$ (hundred pounds). What does the least-squares line forecast for $y =$ miles per gallon?

10. **Basketball: Fouls** Data for this problem are based on information from *STATS Basketball Scoreboard*. It is thought that basketball teams that make too many fouls in a game tend to lose the game even if they otherwise play well. Let x be the number of fouls that were more than (i.e., over and above) the number of fouls made the opposing team made. Let y be the percentage of times the team with the larger number of fouls won the game.

x	0	2	5	6
y	50	45	33	26

Complete parts (a) through (e), given $\Sigma x = 13$, $\Sigma y = 154$, $\Sigma x^2 = 65$, $\Sigma y^2 = 6290$, $\Sigma xy = 411$, and $r \approx -0.988$.

- (f) If a team had $x = 4$ fouls over and above the opposing team, what does the least-squares equation forecast for y ?

11. **Auto Accidents: Age** Data for this problem are based on information taken from the *Wall Street Journal*. Let x be the age in years of a licensed automobile driver. Let y be the percentage of all fatal accidents (for a given age) due to speeding. For example, the first data pair indicates that 36% of all fatal accidents involving 17-year-olds are due to speeding.

x	17	27	37	47	57	67	77
y	36	25	20	12	10	7	5

Complete parts (a) through (e), given $\Sigma x = 329$, $\Sigma y = 115$, $\Sigma x^2 = 18,263$, $\Sigma y^2 = 2639$, $\Sigma xy = 4015$, and $r \approx -0.959$.

- (f) Predict the percentage of all fatal accidents due to speeding for 25-year-olds.

12. **Auto Accidents: Age** Let x be the age of a licensed driver in years. Let y be the percentage of all fatal accidents (for a given age) due to failure to yield the right-of-way. For example, the first data pair states that 5% of all fatal accidents of 37-year-olds are due to failure to yield the right-of-way. The *Wall Street Journal* article referenced in Problem 11 reported the following data:

x	37	47	57	67	77	87
y	5	8	10	16	30	43

Complete parts (a) through (e), given $\Sigma x = 372$, $\Sigma y = 112$, $\Sigma x^2 = 24,814$, $\Sigma y^2 = 3194$, $\Sigma xy = 8254$, and $r \approx -0.943$.

- (f) Predict the percentage of all fatal accidents due to failing to yield the right-of-way for 70-year-olds.

13. **Income: Medical Care** Let x be per capita income in thousands of dollars. Let y be the number of medical doctors per 10,000 residents. Six small cities in Oregon gave the following information about x and y (based on information from *Life in America's Small Cities* by G. S. Thomas, Prometheus Books).

x	8.6	9.3	10.1	8.0	8.3	8.7
y	9.6	18.5	20.9	10.2	11.4	13.1

Complete parts (a) through (e), given $\Sigma x = 53$, $\Sigma y = 83.7$, $\Sigma x^2 = 471.04$, $\Sigma y^2 = 1276.83$, $\Sigma xy = 755.89$, and $r \approx 0.934$.

- (f) Suppose a small city in Oregon has a per capita income of 10 thousand dollars. What is the predicted number of M.D.s per 10,000 residents?

14. **Violent Crimes: Prisons** Does prison really deter violent crime? Let x represent percent change in the rate of violent crime and y represent percent change in the rate of imprisonment in the general U.S. population. For 7 recent years, the following data have been obtained (Source: *The Crime Drop in America*, edited by Blumstein and Wallman, Cambridge University Press).

x	6.1	5.7	3.9	5.2	6.2	6.5	11.1
y	-1.4	-4.1	-7.0	-4.0	3.6	-0.1	-4.4

Complete parts (a) through (e), given $\Sigma x = 44.7$, $\Sigma y = -17.4$, $\Sigma x^2 = 315.85$, $\Sigma y^2 = 116.1$, $\Sigma xy = -107.18$, and $r \approx 0.084$.

- (f) **Critical Thinking** Considering the values of r and r^2 , does it make sense to use the least-squares line for prediction? Explain.

15. **Education: Violent Crime** The following data are based on information from the book *Life in America's Small Cities* (by G. S. Thomas, Prometheus Books). Let x be the percentage of 16- to 19-year-olds not in school and not high school graduates. Let y be the reported violent crimes per 1000 residents. Six small cities in Arkansas (Blytheville, El Dorado, Hot Springs, Jonesboro, Rogers, and Russellville) reported the following information about x and y :

x	24.2	19.0	18.2	14.9	19.0	17.5
y	13.0	4.4	9.3	1.3	0.8	3.6

Complete parts (a) through (e), given $\Sigma x = 112.8$, $\Sigma y = 32.4$, $\Sigma x^2 = 2167.14$, $\Sigma y^2 = 290.14$, $\Sigma xy = 665.03$, and $r \approx 0.764$.

- (f) If the percentage of 16- to 19-year-olds not in school and not graduates reaches 24% in a similar city, what is the predicted rate of violent crimes per 1000 residents?

16. **Research: Patents** The following data are based on information from the *Harvard Business Review* (Vol. 72, No. 1). Let x be the number of different research programs, and let y be the mean number of patents per program. As in any business, a company can spread itself too thin. For example, too many research programs might lead to a decline in overall research productivity. The following data are for a collection of pharmaceutical companies and their research programs:

x	10	12	14	16	18	20
y	1.8	1.7	1.5	1.4	1.0	0.7

Complete parts (a) through (e), given $\Sigma x = 90$, $\Sigma y = 8.1$, $\Sigma x^2 = 1420$, $\Sigma y^2 = 11.83$, $\Sigma xy = 113.8$, and $r \approx -0.973$.

- (f) Suppose a pharmaceutical company has 15 different research programs. What does the least-squares equation forecast for $y =$ mean number of patents per program?

17. **Archaeology: Artifacts** Data for this problem are based on information taken from *Prehistoric New Mexico: Background for Survey* (by D. E. Stuart and R. P. Gauthier, University of New Mexico Press). It is thought that prehistoric Indians did not take their best tools, pottery, and household items when they visited higher elevations for their summer camps. It is hypothesized that archaeological sites tend to lose their cultural identity and specific cultural affiliation as the elevation of the site increases. Let x be the elevation (in thousands of feet) of an archaeological site in the southwestern United States. Let y be the percentage of unidentified artifacts (no specific cultural affiliation) at a given elevation. The following data were obtained for a collection of archaeological sites in New Mexico:

x	5.25	5.75	6.25	6.75	7.25
y	19	13	33	37	62

Complete parts (a) through (e), given $\Sigma x = 31.25$, $\Sigma y = 164$, $\Sigma x^2 = 197.813$, $\Sigma y^2 = 6832$, $\Sigma xy = 1080$, and $r \approx 0.913$.

- (f) At an archaeological site with elevation 6.5 (thousand feet), what does the least-squares equation forecast for $y =$ percentage of culturally unidentified artifacts?



18. **Cricket Chirps: Temperature** Anyone who has been outdoors on a summer evening has probably heard crickets. Did you know that it is possible to use the cricket as a thermometer? Crickets tend to chirp more frequently as temperatures increase. This phenomenon was studied in detail by George W. Pierce, a physics professor at Harvard. In the following data, x is a random variable representing chirps per second and y is a random variable representing temperature ($^{\circ}\text{F}$). These data are also available for download at the Online Study Center.

x	20.0	16.0	19.8	18.4	17.1	15.5	14.7	17.1
y	88.6	71.6	93.3	84.3	80.6	75.2	69.7	82.0

x	15.4	16.2	15.0	17.2	16.0	17.0	14.4
y	69.4	83.3	79.6	82.6	80.6	83.5	76.3

Source: Reprinted by permission of the publisher from *The Songs of Insects* by George W. Pierce, Cambridge, Mass.: Harvard University Press, Copyright © 1948 by the President and Fellows of Harvard College.

Complete parts (a) through (e), given $\Sigma x = 249.8$, $\Sigma y = 1200.6$, $\Sigma x^2 = 4200.56$, $\Sigma y^2 = 96,725.86$, $\Sigma xy = 20,127.47$, and $r \approx 0.835$.

- (f) What is the predicted temperature when $x = 19$ chirps per second?





19. **Expand Your Knowledge: Residual Plot** The least-squares line usually does not go through all the sample data points (x, y) . In fact, for a specified x value from a data pair (x, y) , there is usually a difference between the predicted value and the y value paired with x . This difference is called the *residual*.

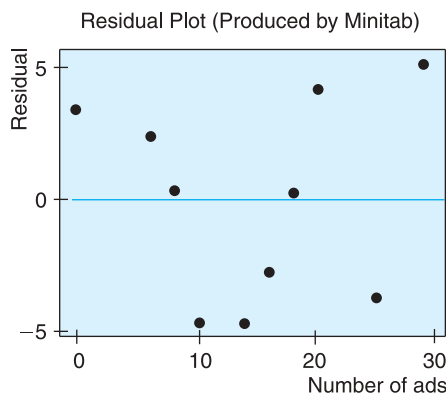
The **residual** is the difference between the y value in a specified data pair (x, y) and the value $\hat{y} = a + bx$ predicted by the least-squares line for the same x .

$$y - \hat{y} \text{ is the residual.}$$

Residual plot

One way to assess how well a least-squares line serves as a model for the data is a **residual plot**. To make a residual plot, we put the x values in order on the horizontal axis and plot the corresponding residuals $y - \hat{y}$ in the vertical direction. Because the mean of the residuals is always zero for a least-squares model, we place a horizontal line at zero. The accompanying figure shows a residual plot for the data of Guided Exercise 4, in which the relationship between the number of ads run per week and the number of cars sold that week was explored. To make the residual plot, first compute all the residuals. Remember that x and y are the given data values, and \hat{y} is computed from the least-squares line $\hat{y} \approx 6.56 + 1.01x$.

Residual				Residual			
x	y	\hat{y}	$y - \hat{y}$	x	y	\hat{y}	$y - \hat{y}$
6	15	12.6	2.4	16	20	22.7	-2.7
20	31	26.8	4.2	28	40	34.8	5.2
0	10	6.6	3.4	18	25	24.7	0.3
14	16	20.7	-4.7	10	12	16.7	-4.7
25	28	31.8	-3.8	8	15	14.6	0.4



- If the least-squares line provides a reasonable model for the data, the pattern of points in the plot will seem random and unstructured about the horizontal line at 0. Is this the case for the residual plot?
- If a point on the residual plot seems far outside the pattern of other points, it might reflect an unusual data point (x, y) , called an *outlier*. Such points may have quite an influence on the least-squares model. Do there appear to be any outliers in the data for the residual plot?



20. **Residual Plot: Miles per Gallon** Consider the data of Problem 9.

- Make a residual plot for the least-squares model.
- Use the residual plot to comment about the appropriateness of the least-squares model for these data. See Problem 19.

21. **Critical Thinking: Exchange x and y in Least-Squares Equation**(a) Suppose you are given the following (x, y) data pairs:

x	1	3	4
y	2	1	6

Show that the least-squares equation for these data is $y = 1.071x + 0.143$ (rounded to three digits after the decimal).

(b) Now suppose you are given these (x, y) data pairs:

x	2	1	6
y	1	3	4

Show that the least-squares equation for these data is $y = 0.357x + 1.595$ (rounded to three digits after the decimal).

- (c) In the data for parts (a) and (b), did we simply exchange the x and y values of each data pair?
- (d) Solve $y = 0.143 + 1.071x$ for x . Do you get the least-squares equation of part (b) with the symbols x and y exchanged?
- (e) In general, suppose we have the least-squares equation $y = a + bx$ for a set of data pairs (x, y) . If we solve this equation for x , will we *necessarily* get the least-squares equation for the set of data pairs (y, x) (with x and y exchanged)? Explain using parts (a) through (d).

22. **Expand Your Knowledge: Logarithmic Transformations, Exponential Growth Model**

There are several extensions of linear regression that apply to exponential growth and power law models. Problems 22–25 will outline some of these extensions. First of all, recall that a variable grows *linearly* over time if it *adds* a fixed increment during each equal time period. *Exponential* growth occurs when a variable is *multiplied* by a fixed number during each time period. This means that exponential growth increases by a fixed multiple or percentage of the previous amount. College algebra can be used to show that if a variable grows exponentially, then its logarithm grows linearly. The exponential growth model is $y = \alpha\beta^x$, where α and β are fixed constants to be estimated from data.

How do we know when we are dealing with exponential growth, and how can we estimate α and β ? Please read on. Populations of living things such as bacteria, locusts, fish, panda bears, and so on, tend to grow (or decline) exponentially. However, these populations can be restricted by outside limitations such as food, space, pollution, disease, hunting, and so on. Suppose we have data pairs (x, y) for which there is reason to believe the scatter plot is not linear, but rather exponential, as described above. This means the increase in y values begins rather slowly but then seems to explode. *Note:* For exponential growth models, we assume all $y > 0$.

Consider the following data, where x = time in hours and y = number of bacteria in a laboratory culture at the end of x hours.

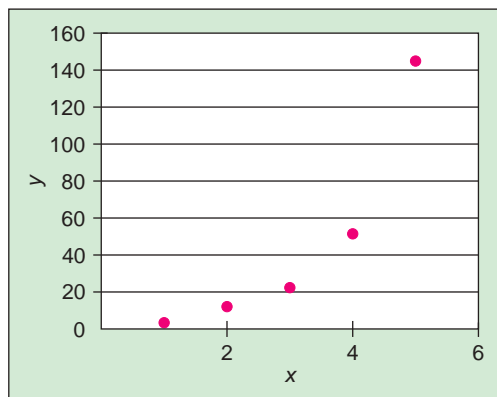
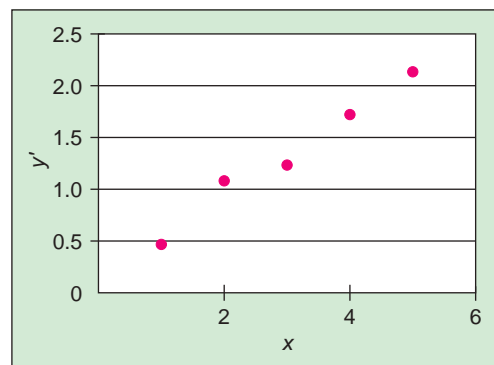
x	1	2	3	4	5
y	3	12	22	51	145

- (a) Look at the Excel graph of the scatter diagram of the (x, y) data pairs. Do you think a straight line will be a good fit to these data? Do the y values seem almost to explode as time goes on?
- (b) Now consider a transformation $y' = \log y$. We are using common logarithms of base 10 (however, natural logarithms of base e would work just as well).

x	1	2	3	4	5
$y' = \log y$	0.477	1.079	1.342	1.748	2.161

Look at the Excel graph of the scatter diagram of the (x, y') data pairs and compare this diagram with the diagram in part (a). Which graph appears to better fit a straight line?

Excel Graphs

Part (a) Model with (x, y) Data PairsPart (b) Model with (x, y') Data Pairs

- (c) Use a calculator with regression keys to verify the linear regression equation for the (x, y) data pairs, $\hat{y} = -50.3 + 32.3x$, with sample correlation coefficient $r = 0.882$.
- (d) Use a calculator with regression keys to verify the linear regression equation for the (x, y') data pairs, $y' = 0.150 + 0.404x$, with sample correlation coefficient $r = 0.994$. The sample correlation coefficient $r = 0.882$ for the (x, y) pairs is not bad. But the sample correlation coefficient $r = 0.994$ for the (x, y') pairs is a lot better!
- (e) The exponential growth model is $y = \alpha\beta^x$. Let us use the results of part (d) to estimate α and β for this strain of laboratory bacteria. The equation $y' = a + bx$ is the same as $\log y = a + bx$. If we raise both sides of this equation to the power 10 and use some college algebra, we get $y = 10^a(10^b)^x$. Thus, $\alpha \approx 10^a$ and $\beta \approx 10^b$. Use these results to approximate α and β and write the exponential growth equation for our strain of bacteria.

Note: The TI-84Plus/TI-83Plus/TI-*n*spire calculators fully support the exponential growth model. Place the original x data in list L1 and the corresponding y data in list L2. Then press **STAT**, followed by **CALC**, and scroll down to option **0: ExpReg**. The output gives values for α , β , and the sample correlation coefficient r .



23. **Expand Your Knowledge: Logarithmic Transformations, Exponential Growth Model** Let x = day of observation and y = number of locusts per square meter during a locust infestation in a region of North Africa.

x	2	3	5	8	10
y	2	3	12	125	630

- (a) Draw a scatter diagram of the (x, y) data pairs. Do you think a straight line will be a good fit to these data? Do the y values almost seem to explode as time goes on?
- (b) Now consider a transformation $y' = \log y$. We are using common logarithms of base 10. Draw a scatter diagram of the (x, y') data pairs and compare this diagram with the diagram of part (a). Which graph appears to better fit a straight line?
- (c) Use a calculator with regression keys to find the linear regression equation for the data pairs (x, y') . What is the correlation coefficient?
- (d) The exponential growth model is $y = \alpha\beta^x$. Estimate α and β and write the exponential growth equation. *Hint:* See Problem 22.



24. *Expand Your Knowledge: Logarithmic Transformations, Power Law Model*

When we take measurements of the same general type, a power law of the form $y = \alpha x^\beta$ often gives an excellent fit to the data. A lot of research has been conducted as to why power laws work so well in business, economics, biology, ecology, medicine, engineering, social science, and so on. Let us just say that if you do not have a good straight-line fit to data pairs (x, y) , and the scatter plot does not rise dramatically (as in exponential growth), then a power law is often a good choice. College algebra can be used to show that power law models become linear when we apply logarithmic transformations to both variables. To see how this is done, please read on. *Note:* For power law models, we assume all $x > 0$ and all $y > 0$.

Suppose we have data pairs (x, y) and we want to find constants α and β such that $y = \alpha x^\beta$ is a good fit to the data. First, make the logarithmic transformations $x' = \log x$ and $y' = \log y$. Next, use the (x', y') data pairs and a calculator with linear regression keys to obtain the least-squares equation $y' = a + bx'$. Note that the equation $y' = a + bx'$ is the same as $\log y = a + b(\log x)$. If we raise both sides of this equation to the power 10 and use some college algebra, we get $y = 10^{a+(bx')}$. In other words, for the power law model, we have $\alpha \approx 10^a$ and $\beta \approx b$.

In the electronic design of a cell phone circuit, the buildup of electric current (Amps) is an important function of time (microseconds). Let $x =$ time in microseconds and let $y =$ Amps built up in the circuit at time x .

x	2	4	6	8	10
y	1.81	2.90	3.20	3.68	4.11

- Make the logarithmic transformations $x' = \log x$ and $y' = \log y$. Then make a scatter plot of the (x', y') values. Does a linear equation seem to be a good fit to this plot?
- Use the (x', y') data points and a calculator with regression keys to find the least-squares equation $y' = a + bx'$. What is the sample correlation coefficient?
- Use the results of part (b) to find estimates for α and β in the power law $y = \alpha x^\beta$. Write the power law giving the relationship between time and Amp buildup.

Note: The TI-84Plus/TI-83Plus/TI-*n*spire calculators fully support the power law model. Place the original x data in list L1 and the corresponding y data in list L2. Then press **STAT**, followed by **CALC**, and scroll down to option **A: PwrReg**. The output gives values for α , β , and the sample correlation coefficient r .



25. *Expand Your Knowledge: Logarithmic Transformations, Power Law Model*

Let $x =$ boiler steam pressure in 100 lb/in.² and let $y =$ critical sheer strength of boiler plate steel joints in tons/in.². We have the following data for a series of factory boilers.

x	4	5	6	8	10
y	3.4	4.2	6.3	10.9	13.3

- Make the logarithmic transformations $x' = \log x$ and $y' = \log y$. Then make a scatter plot of the (x', y') values. Does a linear equation seem to be a good fit to this plot?
- Use the (x', y') data points and a calculator with regression keys to find the least-squares equation $y' = a + bx'$. What is the sample correlation coefficient?
- Use the results of part (b) to find estimates for α and β in the power law $y = \alpha x^\beta$. Write the power equation for the relationship between steam pressure and sheer strength of boiler plate steel. *Hint:* See Problem 24.

SECTION 9.3

Inferences for Correlation and Regression

FOCUS POINTS

- Test the correlation coefficient ρ .
- Use sample data to compute the standard error of estimate S_e .
- Find a confidence interval for the value of y predicted for a specified value of x .
- Test the slope β of the least-squares line.
- Find a confidence interval for the slope β of the least-squares line and interpret its meaning.

Learn more, earn more! We have probably all heard this platitude. The question is whether or not there is some truth in this statement. Do college graduates have an improved chance at a better income? Is there a trend in the general population to support the “learn more, earn more” statement?

Consider the following variables: x = percentage of the population 25 or older with at least four years of college and y = percentage *growth* in per capita income over the past seven years. A random sample of six communities in Ohio gave the information (based on *Life in America's Small Cities* by G. S. Thomas) shown in Table 9-10 on the next page.

If we use what we learned in Sections 9.1 and 9.2, we can compute the sample correlation coefficient r and the least-squares line $\hat{y} = a + bx$ using the data of Table 9-10. However, r is only a *sample* correlation coefficient, and $\hat{y} = a + bx$ is only a “*sample-based*” least-squares line. What if we used *all* possible data pairs (x, y) from *all* U.S. cities, not just six towns in Ohio? If we accomplished this seemingly impossible task, we would have the *population* of all (x, y) pairs.

From this population of (x, y) pairs, we could (in theory) compute the *population correlation coefficient*, which we call ρ (Greek letter rho, pronounced like “row”). We could also compute the least-squares line for the entire population, which we denote as $y = \alpha + \beta x$ using more Greek letters, α (alpha) and β (beta).

Population correlation coefficient ρ

Sample Statistic		Population Parameter
r	→	ρ
a	→	α
b	→	β
$\hat{y} = a + bx$	→	$y = \alpha + \beta x$

Requirements for statistical inference

To make inferences regarding correlation and linear regression, we need to be sure that

- The set (x, y) of ordered pairs is a *random sample* from the population of all possible such (x, y) pairs.
- For each fixed value of x , the y values have a normal distribution. All of the y distributions have the same variance, and, for a given x value, the distribution of y values has a mean that lies on the least-squares line. We also assume that for a fixed y , each x has its own normal distribution. In most cases the results are still accurate if the distributions are simply mound-shaped and symmetric and the y variances are approximately equal.

Requirements for inferences concerning linear regression

We assume these conditions are met for all inferences presented in this section.

Testing the Correlation Coefficient

The first topic we want to study is the statistical significance of the sample correlation coefficient r . To do this, we construct a statistical test of ρ , the population correlation coefficient. The test will be based on the following theorem.

THEOREM 9.1 Let r be the sample correlation coefficient computed using data pairs (x, y) . We use the null hypothesis

$$H_0: x \text{ and } y \text{ have no linear correlation, so } \rho = 0$$

The alternate hypothesis may be

$$H_1: \rho > 0 \quad \text{or} \quad H_1: \rho < 0 \quad \text{or} \quad H_1: \rho \neq 0$$

The conversion of r to a Student's t distribution is

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \quad \text{with } d.f. = n - 2$$

where n is the number of sample data pairs (x, y) ($n \geq 3$).

PROCEDURE

HOW TO TEST THE POPULATION CORRELATION COEFFICIENT ρ

1. Use the *null hypothesis* $H_0: \rho = 0$. In the context of the application, state the *alternate hypothesis* ($\rho > 0$ or $\rho < 0$ or $\rho \neq 0$) and set the *level of significance* α .
2. Obtain a random sample of $n \geq 3$ data pairs (x, y) and compute the *sample test statistic*

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \quad \text{with degrees of freedom } d.f. = n - 2$$

where r is the sample correlation coefficient
 n is the sample size

3. Use a Student's t distribution and the type of test, one-tailed or two-tailed, to find (or estimate) the *P-value* corresponding to the test statistic.
4. *Conclude* the test. If $P\text{-value} \leq \alpha$, then reject H_0 . If $P\text{-value} > \alpha$, then do not reject H_0 .
5. *Interpret your conclusion* in the context of the application.



Problem 13 at the end of this section discusses how sample size might affect the significance of r .

EXAMPLE 6

TESTING ρ

Let's return to our data from Ohio regarding the percentage of the population with at least four years of college and the percentage of growth in per capita income (Table 9-10). We'll develop a test for the population correlation coefficient ρ .

SOLUTION: First, we compute the sample correlation coefficient r . Using a calculator, statistical software, or a “by-hand” calculation from Section 9.1, we find

$$r \approx 0.887$$

Now we test the population correlation coefficient ρ . Remember that x represents percentage college graduates and y represents percentage salary increases in

TABLE 9-10 Education and Income Growth Percentages

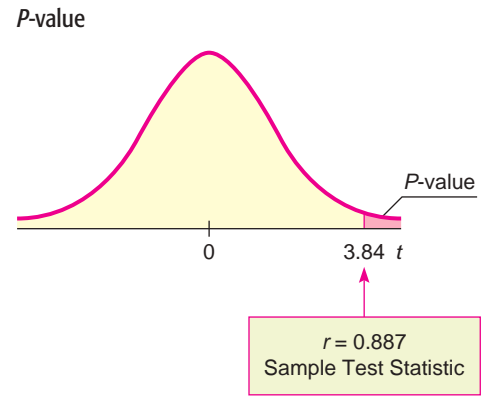
x	9.9	11.4	8.1	14.7	8.5	12.6
y	37.1	43.0	33.4	47.1	26.5	40.2

TABLE 9-11 Excerpt from Student's *t* Distribution

✓one-tail area	0.010	0.005
two-tail area	0.020	0.010
<i>d.f.</i> = 4	3.747	4.604

↑
Sample $t = 3.84$

FIGURE 9-15



the general population. We suspect the population correlation is positive, $\rho > 0$. Let's use a 1% level of significance:

$$H_0: \rho = 0 \text{ (no linear correlation)}$$

$$H_1: \rho > 0 \text{ (positive linear correlation)}$$

Convert the sample test statistic $r = 0.887$ to t using $n = 6$.

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.887\sqrt{6-2}}{\sqrt{1-0.887^2}} \approx 3.84 \quad \text{with } d.f. = n - 2 = 6 - 2 = 4$$

The P -value for the sample test statistic $t = 3.84$ is shown in Figure 9-15. Since we have a right-tailed test, we use the one-tail area in the Student's t distribution (Table 6 of Appendix II).

From Table 9-11, we see that

$$0.005 < P\text{-value} < 0.010$$

Since the interval containing the P -value is less than the level of significance $\alpha = 0.01$, we reject H_0 and conclude that the population correlation coefficient between x and y is positive. Technology gives P -value ≈ 0.0092 .



Caution: Although we have shown that x and y are positively correlated, we have not shown that an increase in education *causes* an increase in earnings.

GUIDED EXERCISE 6

Testing ρ

A medical research team is studying the effect of a new drug on red blood cells. Let x be a random variable representing milligrams of the drug given to a patient. Let y be a random variable representing red blood cells per cubic milliliter of whole blood. A random sample of $n = 7$ volunteer patients gave the following results.

x	9.2	10.1	9.0	12.5	8.8	9.1	9.5
y	5.0	4.8	4.5	5.7	5.1	4.6	4.2

Use a calculator to verify that $r \approx 0.689$. Then use a 1% level of significance to test the claim that $\rho \neq 0$.

Continued

GUIDED EXERCISE 6 *continued*

(a) State the null and alternate hypotheses. What is the level of significance α ?

→ $H_0: \rho = 0; H_1: \rho \neq 0; \alpha = 0.01$

(b) Compute the sample test statistic.

→ $t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \approx \frac{0.689\sqrt{7-2}}{\sqrt{1-0.689^2}} \approx \frac{1.5406}{0.7248} \approx 2.126$

(c) Use the Student's t distribution, Table 6 of Appendix II, to estimate the P -value.

→ $d.f. = n - 2 = 7 - 2 = 5$; two-tailed test

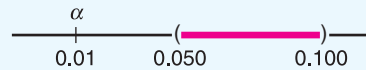
✓ two-tail area	0.100	0.050
$d.f. = 5$	2.015	2.571

↑
Sample $t = 2.126$

$0.050 < P\text{-value} < 0.100$

(d) Do we reject or fail to reject H_0 ?

→ Since the interval containing the P -value lies to the right of $\alpha = 0.01$, we do not reject H_0 . Technology gives $P\text{-value} \approx 0.0866$.



(e) *Interpret* the conclusion in the context of the application.

→ At the 1% level of significance, the evidence is not strong enough to indicate any correlation between the amount of drug administered and the red blood cell count.

Standard Error of Estimate

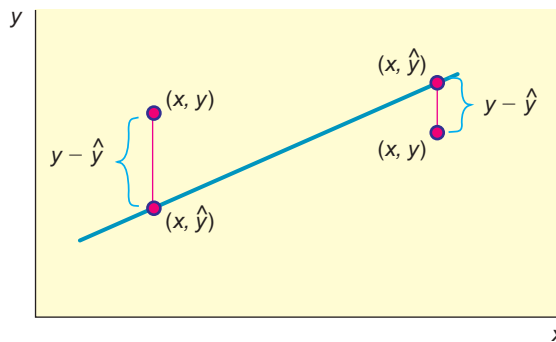
Sometimes a scatter diagram clearly indicates the existence of a linear relationship between x and y , but it can happen that the points are widely scattered about the least-squares line. We need a method (besides just looking) for measuring the spread of a set of points about the least-squares line. There are three common methods of measuring the spread. One method uses the *standard error of estimate*. The others use the *coefficient of correlation* and the *coefficient of determination*.

For the standard error of estimate, we use a measure of spread that is in some ways like the standard deviation of measurements of a single variable. Let

$$\hat{y} = a + bx$$

FIGURE 9-16

The Distance Between Points (x, y) and (x, \hat{y})



be the predicted value of y from the least-squares line. Then $y - \hat{y}$ is the difference between the y value of the *data point* (x, y) shown on the scatter diagram (Figure 9-16) and the \hat{y} value of the point on the *least-squares line* with the same x value. The quantity $y - \hat{y}$ is known as the *residual*. To avoid the difficulty of having some positive and some negative values, we square the quantity $(y - \hat{y})$. Then we sum the squares and, for technical reasons, divide this sum by $n - 2$. Finally, we take the square root to obtain the *standard error of estimate*, denoted by S_e .

Residual

Standard error of estimate S_e

$$\text{Standard error of estimate} = S_e = \sqrt{\frac{\sum (y - \hat{y})^2}{n - 2}} \quad (7)$$

where $\hat{y} = a + bx$ and $n \geq 3$.

Note: To compute the standard error of estimate, we require that there be at least three points on the scatter diagram. If we had only two points, the line would be a perfect fit, since two points determine a line. In such a case, there would be no need to compute S_e .

The nearer the scatter points lie to the least-squares line, the smaller S_e will be. In fact, if $S_e = 0$, it follows that each $y - \hat{y}$ is also zero. This means that all the scatter points lie on the least-squares line if $S_e = 0$. The larger S_e becomes, the more scattered the points are.

The formula for the standard error of estimate is reminiscent of the formula for the standard deviation, which is also a measure of dispersion. However, the standard deviation involves differences of data values from a mean, whereas the standard error of estimate involves the differences between experimental and predicted y values for a given x (i.e., $y - \hat{y}$).

The actual computation of S_e using Equation (7) is quite long because the formula requires us to use the least-squares line equation to compute a predicted value \hat{y} for each x value in the data pairs. There is a computational formula that we strongly recommend you use. However, as with all the computation formulas, be careful about rounding. This formula is sensitive to rounding, and you should carry as many digits as seem reasonable for your problem. Answers will vary, depending on the rounding used. We give the formula here and follow it with an example of its use.

PROCEDURE

HOW TO FIND THE STANDARD ERROR OF ESTIMATE S_e

1. Obtain a random sample of $n \geq 3$ data pairs (x, y) .
2. Use the procedures of Section 9.2 to find a and b from the sample least-squares line $\hat{y} = a + bx$.
3. The standard error of estimate is

$$S_e = \sqrt{\frac{\sum y^2 - a\sum y - b\sum xy}{n - 2}} \quad (8)$$

Computation formula for S_e

With a considerable amount of algebra, Equations (7) and (8) can be shown to be mathematically equivalent. Equation (7) shows the strong similarity between the standard error of estimate and the standard deviation. Equation (8) is a shortcut calculation formula because it involves few subtractions. The sums $\sum x$, $\sum y$, $\sum x^2$, $\sum y^2$, and $\sum xy$ are provided directly on most calculators that support two-variable statistics.

In the next example, we show you how to compute the standard error of estimate using the computation formula.

EXAMPLE 7

LEAST-SQUARES LINE AND S_e

June and Jim are partners in the chemistry lab. Their assignment is to determine how much copper sulfate (CuSO_4) will dissolve in water at 10, 20, 30, 40, 50, 60, and 70°C. Their lab results are shown in Table 9-12, where y is the weight in grams of copper sulfate that will dissolve in 100 grams of water at $x^\circ\text{C}$.

Sketch a scatter diagram, find the equation of the least-squares line, and compute S_e .

Photo Researchers



SOLUTION: Figure 9-17 includes a scatter diagram for the data of Table 9-12. To find the equation of the least-squares line and the value of S_e , we set up a computational table (Table 9-13).

$$\begin{aligned} \bar{x} &= \frac{\Sigma x}{n} = \frac{280}{7} = 40 & \text{and} & \quad \bar{y} = \frac{\Sigma y}{n} = \frac{213}{7} \approx 30.429 \\ b &= \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{n\Sigma x^2 - (\Sigma x)^2} = \frac{7(9940) - (280)(213)}{7(14,000) - (280)^2} = \frac{9940}{19,600} \approx 0.50714 \\ a &= \bar{y} - b\bar{x} \approx 30.429 - 0.507(40) \approx 10.149 \end{aligned}$$

The equation of the least-squares line is

$$\begin{aligned} \hat{y} &= a + bx \\ \hat{y} &\approx 10.14 + 0.51x \end{aligned}$$

The graph of the least-squares line is shown in Figure 9-17. Notice that it passes through the point $(\bar{x}, \bar{y}) = (40, 30.4)$. Another point on the line can be found by using $x = 15$ in the equation of the line $\hat{y} = 10.14 + 0.51x$. When we use 15 in place of x , we obtain $\hat{y} = 10.14 + 0.51(15) = 17.8$. The point $(15, 17.8)$ is the other point we used to graph the least-squares line in Figure 9-17.

The standard error of estimate is computed using the computational formula

$$\begin{aligned} S_e &= \sqrt{\frac{\Sigma y^2 - a\Sigma y - b\Sigma xy}{n - 2}} \\ &\approx \sqrt{\frac{7229 - 10.149(213) - 0.507(9940)}{7 - 2}} \approx \sqrt{\frac{27.683}{5}} \approx 2.35 \end{aligned}$$

Note: This formula is very sensitive to rounded values of a and b .

TABLE 9-12 Lab Results ($x = ^\circ\text{C}$, $y = \text{amount of CuSO}_4$)

x	y
10	17
20	21
30	25
40	28
50	33
60	40
70	49

FIGURE 9-17

Scatter Diagram and Least-Squares Line for Chemistry Experiment

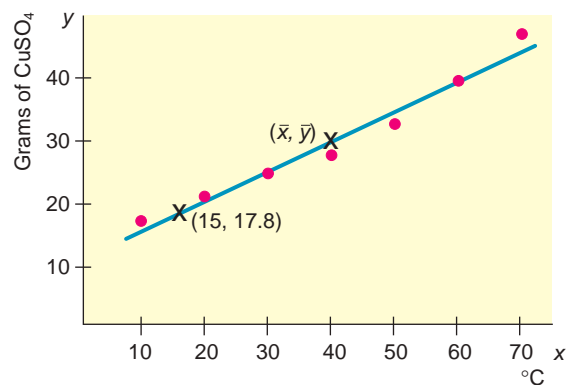


TABLE 9-13 Computational Table

x	y	x^2	y^2	xy
10	17	100	289	170
20	21	400	441	420
30	25	900	625	750
40	28	1600	784	1120
50	33	2500	1089	1650
60	40	3600	1600	2400
70	49	4900	2401	3430
$\Sigma x = 280$	$\Sigma y = 213$	$\Sigma x^2 = 14,000$	$\Sigma y^2 = 7229$	$\Sigma xy = 9940$

TECH NOTES

Although many calculators that support two-variable statistics and linear regression do not provide the value of the standard error of estimate S_e directly, they do provide the sums required for the calculation of S_e . The TI-84Plus/TI-83Plus/TI-*n*spire, Excel 2007, and Minitab all provide the value of S_e .

TI-84Plus/TI-83Plus/TI-*n*spire (with TI-84Plus keypad) The value for S_e is given as s under STAT, TEST, option E: LinRegTTest.

Excel 2007 Click the **Insert Function** $\left(\frac{f_x}{\square}\right)$. In the dialogue box, use **Statistical** for the category, and select the function **STEYX**.

Minitab Use the menu choices **Stat** \blacktriangleright **Regression** \blacktriangleright **Regression**. The value for S_e is given as s in the display.

Confidence Intervals for y

The least-squares line gives us a predicted value \hat{y} for a specified x value. However, we used sample data to get the equation of the line. The line derived from the population of all data pairs is likely to have a slightly different slope, which we designate by the symbol β for population slope, and a slightly different y intercept, which we designate by the symbol α for population intercept. In addition, there is some random error ε , so the true y value is

$$y = \alpha + \beta x + \varepsilon$$

Because of the random variable ε , for each x value there is a corresponding distribution of y values. The methods of linear regression were developed so that the distribution of y values for a given x is centered on the population regression line. Furthermore, the distributions of y values corresponding to each x value all have the same standard deviation, estimated by the standard error of estimate S_e .

Using all this background, the theory tells us that for a specific x , a c confidence interval for y is given by the next procedure.

Population slope β Confidence interval for predicted y

PROCEDURE

HOW TO FIND A CONFIDENCE INTERVAL FOR A PREDICTED y FROM THE LEAST-SQUARES LINE

1. Obtain a random sample of $n \geq 3$ data pairs (x, y) .
2. Use the procedure of Section 9.2 to find the least-squares line $\hat{y} = a + bx$. You also need to find \bar{x} from the sample data and the standard error of estimate S_e using Equation (8) of this section.
3. The c confidence interval for y for a specified value of x is

$$\hat{y} - E < y < \hat{y} + E$$

where

$$E = t_c S_e \sqrt{1 + \frac{1}{n} + \frac{n(x - \bar{x})^2}{n \sum x^2 - (\sum x)^2}}$$

$\hat{y} = a + bx$ is the predicted value of y from the least-squares line for a specified x value

c = confidence level ($0 < c < 1$)

n = number of data pairs ($n \geq 3$)

t_c = critical value from Student's t distribution for c confidence level using $d.f. = n - 2$

S_e = standard error of estimate

The formulas involved in the computation of a c confidence interval look complicated. However, they involve quantities we have already computed or values we can easily look up in tables. The next example illustrates this point.

EXAMPLE 8 CONFIDENCE INTERVAL FOR PREDICTION

Using the data of Table 9-13 of Example 7, find a 95% confidence interval for the amount of copper sulfate that will dissolve in 100 grams of water at 45°C.

SOLUTION: First, we need to find \hat{y} for $x = 45^\circ\text{C}$. We use the equation of the least-squares line that we found in Example 7.

$$\begin{aligned}\hat{y} &\approx 10.14 + 0.51x && \text{from Example 7} \\ \hat{y} &\approx 10.14 + 0.51(45) && \text{use 45 in place of } x \\ \hat{y} &\approx 33\end{aligned}$$

A 95% confidence interval for y is then

$$\begin{aligned}\hat{y} - E &< y < \hat{y} + E \\ 33 - E &< y < 33 + E\end{aligned}$$

$$\text{where } E = t_c S_e \sqrt{1 + \frac{1}{n} + \frac{n(x - \bar{x})^2}{n\sum x^2 - (\sum x)^2}}$$

From Example 7, we have $n = 7$, $\sum x = 280$, $\sum x^2 = 14,000$, $\bar{x} = 40$, and $S_e \approx 2.35$. Using $n - 2 = 7 - 2 = 5$ degrees of freedom, we find from Table 6 of Appendix II that $t_{0.95} = 2.571$.

$$\begin{aligned}E &\approx (2.571)(2.35)\sqrt{1 + \frac{1}{7} + \frac{7(45 - 40)^2}{7(14,000) - (280)^2}} \\ &\approx (2.571)(2.35)\sqrt{1.15179} \approx 6.5\end{aligned}$$

A 95% confidence interval for y is

$$\begin{aligned}33 - 6.5 &\leq y \leq 33 + 6.5 \\ 26.5 &\leq y \leq 39.5\end{aligned}$$

This means we are 95% sure that the interval between 26.5 grams and 39.5 grams is one that contains the predicted amount of copper sulfate that will dissolve in 100 grams of water at 45°C. The interval is fairly wide but would decrease with more sample data.

GUIDED EXERCISE 7

Confidence interval for prediction

Let's use the data of Example 7 to compute a 95% confidence interval for $y =$ amount of copper sulfate that will dissolve at $x = 15^\circ\text{C}$.

- (a) From Example 7, we have

$$\hat{y} \approx 10.14 + 0.51x$$

Evaluate \hat{y} for $x = 15$.

$$\begin{aligned}\Rightarrow \hat{y} &\approx 10.14 + 0.51x \\ &\approx 10.14 + 0.51(15) \\ &\approx 17.8\end{aligned}$$

- (b) The bound E on the error of estimate is

$$E = t_c S_e \sqrt{1 + \frac{1}{n} + \frac{n(x - \bar{x})^2}{n\sum x^2 - (\sum x)^2}}$$

From Example 7, we know that $S_e \approx 2.35$, $\sum x = 280$, $\sum x^2 = 14,000$, $\bar{x} = 40$, and $n = 7$. Find $t_{0.95}$ and compute E .

$$\begin{aligned}\Rightarrow t_{0.95} &= 2.571 \text{ for } d.f. = n - 2 = 5 \\ E &\approx (2.571)(2.35)\sqrt{1 + \frac{1}{7} + \frac{7(15 - 40)^2}{7(14,000) - (280)^2}} \\ &\approx (2.571)(2.35)\sqrt{1.366071} \approx 7.1\end{aligned}$$

Continued

GUIDED EXERCISE 7 *continued*(c) Find a 95% confidence interval for y .

$$\hat{y} - E \leq y \leq \hat{y} + E$$



The confidence interval is

$$17.8 - 7.1 \leq y \leq 17.8 + 7.1$$

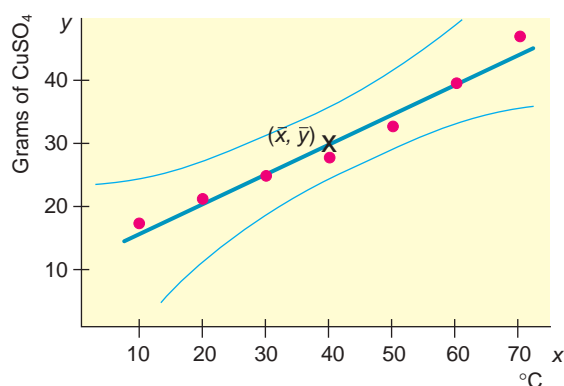
$$10.7 \leq y \leq 24.9$$

As we compare the results of Guided Exercise 7 and Example 8, we notice that the 95% confidence interval of y values for $x = 15^\circ\text{C}$ is 7.1 units above and below the least-squares line, while the 95% confidence interval of y values for $x = 45^\circ\text{C}$ is only 6.5 units above and below the least-squares line. This comparison reflects the general property that confidence intervals for y are narrower the nearer we are to the mean \bar{x} of the x values. As we move near the extremes of the x distribution, the confidence intervals for y become wider. This is another reason that we should not try to use the least-squares line to predict y values for x values beyond the data extremes of the sample x distribution.

If we were to compute a 95% confidence interval for all x values in the range of the sample x values, the *confidence interval band* would curve away from the least-squares line, as shown in Figure 9-18.

Confidence prediction band

FIGURE 9-18

95% Confidence Band for Predicted Values \hat{y} 

TECH NOTES

Minitab provides confidence intervals for predictions. Use the menu selection **Stat** ► **Regression** ► **Regression**. Under **Options**, enter the observed x value and set the confidence level. In the output, the confidence interval for predictions is designated by %PI.

Inferences about the Slope β

Recall that $\hat{y} = a + bx$ is the sample-based least-squares line and that $y = \alpha + \beta x$ is the population-based least-squares line computed (in theory) from the population of all (x, y) data pairs. In many real-world applications, the slope β is very important because β measures the rate at which y changes per unit change in x . Our next topic is to develop statistical tests and confidence intervals for β . Our work is based on the following theorem.

THEOREM 9.2 Let b be the slope of the sample least-squares line $\hat{y} = a + bx$ computed from a random sample of $n \geq 3$ data pairs (x, y) . Let β be the slope of the population least-squares line $y = \alpha + \beta x$, which is in theory computed from the population of all (x, y) data pairs. Let S_e be the standard error of estimate computed from the sample. Then

$$t = \frac{b - \beta}{S_e / \sqrt{\Sigma x^2 - \frac{1}{n} (\Sigma x)^2}}$$

has a Student's t distribution with degrees of freedom $d.f. = n - 2$.

COMMENT The expression $S_e / \sqrt{\Sigma x^2 - \frac{1}{n} (\Sigma x)^2}$ is called the *standard error* for b .

Using this theorem, we can construct procedures for statistical tests and confidence intervals for β .

PROCEDURE

HOW TO TEST β AND FIND A CONFIDENCE INTERVAL FOR β

Requirements

Obtain a random sample of $n \geq 3$ data pairs (x, y) . Use the procedure of Section 9.2 to find b , the slope of the sample least-squares line. Use Equation (8) of this section to find S_e , the standard error of estimate.

Procedure

For a statistical test of β

1. Use the *null hypothesis* $H_0: \beta = 0$. Use an *alternate hypothesis* H_1 appropriate to your application ($\beta > 0$ or $\beta < 0$ or $\beta \neq 0$). Set the level of significance α .
2. Use the null hypothesis $H_0: \beta = 0$ and the values of S_e , n , Σx , Σx^2 , and b to compute the *sample test statistic*.

$$t = \frac{b}{S_e} \sqrt{\Sigma x^2 - \frac{1}{n} (\Sigma x)^2} \quad \text{with } d.f. = n - 2$$

3. Use a Student's t distribution and the type of test, one-tailed or two-tailed, to find (or estimate) the *P-value* corresponding to the test statistic.
4. *Conclude* the test. If $P\text{-value} \leq \alpha$, then reject H_0 . If $P\text{-value} > \alpha$, then do not reject H_0 .
5. *Interpret your conclusion* in the context of the application.

To find a confidence interval for β

$$b - E < \beta < b + E$$

$$\text{where } E = \frac{t_c S_e}{\sqrt{\Sigma x^2 - \frac{1}{n} (\Sigma x)^2}}$$

c = confidence level ($0 < c < 1$)

n = number of data pairs (x, y) , $n \geq 3$

t_c = Student's t distribution critical value for confidence level c and $d.f. = n - 2$

S_e = standard error of estimate

Hypothesis test for β

Confidence interval for β

EXAMPLE 9

TESTING β AND FINDING A CONFIDENCE INTERVAL FOR β

Plate tectonics and the spread of the ocean floor are very important topics in modern studies of earthquakes and earth science in general. A random sample of islands in the Indian Ocean gave the following information.

x = age of volcanic island in the Indian Ocean (units in 10^6 years)
 y = distance of the island from the center of the midoceanic ridge (units in 100 kilometers)

x	120	83	60	50	35	30	20	17
y	30	16	15.5	14.5	22	18	12	0

Source: From King, Cuchaine A. M. *Physical Geography*. Oxford: Basil Blackwell, 1980, pp. 77–86 and 196–206. Reprinted by permission of the publisher.



Photo Researchers

(a) Starting from raw data values (x, y) , the first step is simple but tedious. In short, you may verify (if you wish) that

$$\Sigma x = 415, \Sigma y = 128, \Sigma x^2 = 30,203, \Sigma y^2 = 2558.5,$$

$$\Sigma xy = 8133, \bar{x} = 51.875, \text{ and } \bar{y} = 16$$

(b) The next step is to compute b , a , and S_e . Using a calculator, statistical software, or the formulas, we get

$$b \approx 0.1721 \quad \text{and} \quad a \approx 7.072$$

Since $n = 8$, we get

$$S_e = \sqrt{\frac{\Sigma y^2 - a\Sigma y - b\Sigma xy}{n - 2}}$$

$$\approx \sqrt{\frac{2558.5 - 7.072(128) - 0.1721(8133)}{8 - 2}} \approx 6.50$$

(c) Use an $\alpha = 5\%$ level of significance to test the claim that β is positive.

SOLUTION: $\alpha = 0.05$; $H_0: \beta = 0$; $H_1: \beta > 0$. The sample test statistic is

$$t = \frac{b}{S_e} \sqrt{\Sigma x^2 - \frac{1}{n} (\Sigma x)^2} \approx \frac{0.1721}{6.50} \sqrt{30,203 - \frac{(415)^2}{8}} \approx 2.466$$

with $d.f. = n - 2 = 8 - 2 = 6$.

We use the Student's t distribution (Table 6 of Appendix II) to find an interval containing the P -value. The test is a one-tailed test. Technology gives P -value ≈ 0.0244 .

TABLE 9-14 Excerpt from Table 6, Appendix II

✓ one-tail area	0.025	0.010
$d.f. = 6$	2.447	3.143
	↑	
	Sample $t = 7.563$	

$$0.010 < P\text{-value} < 0.025$$



Since the interval containing the P -value is less than $\alpha = 0.05$, we reject H_0 and conclude that, at the 5% level of significance, the slope is positive.

(d) Find a 75% confidence interval for β .

SOLUTION: For $c = 0.75$ and $d.f. = n - 2 = 8 - 2 = 6$, the critical value $t_c = 1.273$. The margin of error E for the confidence interval is

$$E = \frac{t_c S_e}{\sqrt{\Sigma x^2 - \frac{1}{n} (\Sigma x)^2}} \approx \frac{1.273(6.50)}{\sqrt{30,203 - \frac{(415)^2}{8}}} \approx 0.0888$$

Using $b \approx 0.17$, a 75% confidence interval for β is

$$b - E < \beta < b + E$$

$$0.17 - 0.09 < \beta < 0.17 + 0.09$$

$$0.08 < \beta < 0.26$$

- (e) **Interpretation** What does the confidence interval mean? Recall the units involved (x in 10^6 years and y in 100 kilometers). It appears that, in this part of the world, we can be 75% confident that we have an interval showing that the ocean floor is moving at a rate of between 8 mm and 26 mm per year.

GUIDED EXERCISE 8

Inference for β

How fast do puppies grow? That depends on the puppy. How about male wolf pups in the Helsinki Zoo (Finland)? Let x = age in weeks and y = weight in kilograms for a random sample of male wolf pups. The following data are based on the article “Studies of the Wolf in Finland *Canis lupus L*” (*Ann. Zool. Fenn.*, Vol. 2, pp. 215–259) by E. Pulliainen, University of Helsinki.

x	8	10	14	20	28	40	45
y	7	13	17	23	30	34	35

BAUER, ERWIN & PEGGY/Animals Animals-Earth Scenes-All rights reserved.



$\Sigma x = 165, \Sigma y = 159, \Sigma y^2 = 5169, \Sigma x^2 = 4317, \Sigma xy = 4659$

- (a) Verify the following values.
 $\bar{x} \approx 23.571, \bar{y} \approx 22.714,$
 $b \approx 0.7120, a \approx 5.932,$
 $S_e \approx 3.368$
- (b) Use a 1% level of significance to test the claim that $\beta \neq 0$, and **interpret** the results in the context of this application.

➔ Use the formulas for \bar{x}, \bar{y}, b, a and S_e or find the results directly using your calculator or computer software.

➔ $\alpha = 0.01; H_0: \beta = 0; H_1: \beta \neq 0$
 Convert $b \approx 0.7120$ to a t value.

$$t = \frac{b}{S_e} \sqrt{\Sigma x^2 - \frac{1}{n} (\Sigma x)^2}$$

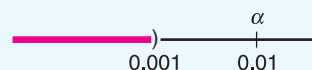
$$\approx \frac{0.7120}{3.368} \sqrt{5169 - \frac{(165)^2}{7}} \approx 7.563$$

From Table 6, Appendix II, for a two-tailed test with $d.f. = n - 2 = 7 - 2 = 5$,

✓ two-tail area	0.001
$d.f. = 5$	6.869

↑
Sample $t = 7.563$

Noting that areas decrease as t values increase, we have $0.001 > P$ -value. Technology gives P -value ≈ 0.0006 .



Since the P -value is less than $\alpha = 0.01$, we reject H_0 and conclude that the population slope β is not zero.

Continued

GUIDED EXERCISE 8 *continued*

- (c) Compute an 80% confidence interval for β and *interpret* the results in the context of this application.

➔ $d.f. = 5$. For an 80% confidence interval, the critical value $t_c = 1.476$. The confidence interval is

$$b - E < \beta < b + E$$

where $b = 0.712$ and

$$E = \frac{t_c S_c}{\sqrt{\sum x^2 - \frac{(\sum x)^2}{n}}} \approx \frac{1.476(3.368)}{\sqrt{5169 - \frac{(165)^2}{7}}} \approx 0.139$$

The interval is from 0.57 kg to 0.85 kg. We can be 80% confident that the interval computed is one that contains β . For each week's change in age, the weight change is between 0.57 kg and 0.85 kg.

Computation Hints for Sample Test Statistic t Used in Testing ρ and Testing β

In Section 9.2 we saw that the values for the sample correlation coefficient r and slope b of the least-squares line are related by the formula

$$b = r \left(\frac{s_y}{s_x} \right)$$

Using this relationship and some algebra, it can be shown that

For the same sample, the sample correlation coefficient r and the slope b of the least-squares line have the same sample test statistic t , with $d.f. = n - 2$, where n is the number of data pairs.

Consequently, when doing calculations or using results from technology, we can use the following strategies.

- Calculations "by hand":** Find the sample test statistic t corresponding to r . The sample test statistic t corresponding to b is the same.
- Using computer results:** Most computer-based statistical packages provide the sample test statistic t corresponding to b . The sample test statistic t corresponding to r is the same.

TECH NOTES

The sample test statistic t corresponding to the sample correlation coefficient r is the same as the t value corresponding to b , the slope of the least-squares line (see Problem 14 at the end of this section). Consequently, the two tests $H_0: \rho = 0$ and $H_0: \beta = 0$ (with similar corresponding alternate hypotheses) have the same conclusions. The TI-84Plus/TI-83Plus/TI-*n*spire calculators use this fact explicitly. Minitab and Excel 2007 show t and the two-tailed P -value for the slope b of the least-squares line. Excel also shows confidence intervals for β . The displays show data from Guided Exercise 8 regarding the age and weight of wolf pups.

TI-84Plus/TI-83Plus/TI-*n*spire (with TI-84Plus keypad) Under STAT, select TEST and use option E:LinRegTTest.



Problem 14 discusses the fact that for the same data, the values of the sample test statistics for r and b are equal.

```

LinRegTTest
y=a+bx
B≠0 and p≠0
↑b=.7120
s=3.3676
r2=.9196
r=.9590

```

```

LinRegTTest
y=a+bx
B≠0 and p≠0
t=7.5632
p=6.4075E-4
df=5.0000
↓a=5.9317

```

Note that the value of S_e is given as s .

Excel 2007 On the home screen, click the **Data** tab. Select **Data Analysis** in the Analysis group. In the dialogue box, select **Regression**. Widen columns of the output as necessary to see all the results.

Regression Statistics	
Multiple R	0.958966516
R Square	0.919616778
Adjusted R Square	0.903540133
Standard Error	3.367628886 ← Value of S_e
Observations	7

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	5.931681179	2.558126184	2.318760198	0.068158803	-0.644180779	12.50754314
X Variable 1	0.711989283	0.094138596	7.563202697	0.000640746	0.469998714	0.953979853

↑
 b
↑
for two-tailed test

Minitab Use the menu selection **Stat** ► **Regression** ► **Regression**. The value of S_e is S ; P is the P -value of a two-tailed test. For a one-tailed test, divide the P -value by 2.

Regression Analysis

The regression equation is

$$y = 5.93 + 0.712 x$$

Predictor	Coef	StDev	T	P
Constant	5.932	2.558	2.32	0.068
x	0.71199	0.09414	7.56	0.001
S = 3.368		R-Sq = 92.0%	R-Sq(adj) = 90.4%	

VIEWPOINT

Hawaiian Island Hopping!

Suppose you want to go camping in Hawaii. Yes! Hawaii has both state and federal parks where you can enjoy camping on the beach or in the mountains. However, you will probably need to rent a car to get to the different campgrounds. How much will the car rental cost? That depends on the islands you visit. For car rental data and regression statistics you can compute regarding costs on different Hawaiian Islands, visit the Brase/Brase statistics site at <http://www.cengage.com/statistics/brase> and find the link to Hawaiian Islands.

SECTION 9.3
PROBLEMS

- Statistical Literacy** What is the symbol used for the population correlation coefficient?
- Statistical Literacy** What is the symbol used for the slope of the population least-squares line?
- Statistical Literacy** For a fixed confidence level, how does the length of the confidence interval for predicted values of y change as the corresponding x values become further away from \bar{x} ?
- Statistical Literacy** How does the t value for the sample correlation coefficient r compare to the t value for the corresponding slope b of the sample least-squares line?

Using Computer Printouts Problems 5 and 6 use the following information. Prehistoric pottery vessels are usually found as sherds (broken pieces) and are carefully reconstructed if enough sherds can be found. Information taken from *Mimbres Mogollon Archaeology* by A. I. Woosley and A. J. McIntyre (University of New Mexico Press) provides data relating x = body diameter in centimeters and y = height in centimeters of prehistoric vessels reconstructed from sherds found at a prehistoric site. The following Minitab printout provides an analysis of the data.

Predictor	Coef	SE Coef	T	P
Constant	-0.223	2.429	-0.09	0.929
Diameter	0.7848	0.1471	5.33	0.001

S = 4.07980 R-Sq = 80.3%

- Critical Thinking: Using Information from a Computer Display to Test for Significance** Refer to the Minitab printout regarding prehistoric pottery.
 - Minitab calls the explanatory variable the predictor variable. Which is the predictor variable, the diameter of the pot or the height?
 - For the least-squares line $\hat{y} = a + bx$, what is the value of the constant a ? What is the value of the slope b ? (*Note:* The slope is the coefficient of the predictor variable.) Write the equation of the least-squares line.
 - The P -value for a two-tailed test corresponding to each coefficient is listed under “P.” The t value corresponding to the coefficient is listed under “T.” What is the P -value of the slope? What are the hypotheses for a two-tailed test of $\beta = 0$? Based on the P -value in the printout, do we reject or fail to reject the null hypothesis for $\alpha = 0.01$?
 - Recall that the t value and resulting P -value of the slope b equal the t value and resulting P -value of the corresponding sample correlation coefficient r . To find the value of the sample correlation coefficient r , take the square root of the R-Sq value shown in the display. What is the value of r ? Consider a two-tailed test for ρ . Based on the P -value shown in the Minitab display, is the correlation coefficient significant at the 1% level of significance?
- Critical Thinking: Using Information from a Computer Display to Find a Confidence Interval** Refer to the Minitab printout regarding prehistoric pottery.
 - The standard error S_e of the linear regression model is given in the printout as “S.” What is the value of S_e ?
 - The standard error of the coefficient of the predictor variable is found under “SE Coef.” Recall that the standard error for b is $S_e \sqrt{\sum x^2 - \frac{1}{n} (\sum x)^2}$. From the Minitab display, what is the value of the standard error for the slope b ?

- (c) The formula for the margin of error E for a $c\%$ confidence interval for the slope β can be written as $E = t_c(\text{SE Coef})$. The Minitab display is based on $n = 9$ data pairs. Find the critical value t_c for a 95% confidence interval in Table 6 of Appendix II. Then find a 95% confidence interval for the population slope β .

In Problems 7–12, parts (a) and (b) relate to testing ρ . Part (c) requests the value of S_e . Parts (d) and (e) relate to confidence intervals for prediction. Parts (f) and (g) relate to testing β and finding confidence intervals for β .

Answers may vary due to rounding.

7. **Basketball: Free Throws and Field Goals** Let x be a random variable that represents the percentage of successful free throws a professional basketball player makes in a season. Let y be a random variable that represents the percentage of successful field goals a professional basketball player makes in a season. A random sample of $n = 6$ professional basketball players gave the following information (Reference: *The Official NBA Basketball Encyclopedia*, Villard Books).

x	67	65	75	86	73	73
y	44	42	48	51	44	51

- (a) Verify that $\Sigma x = 439$, $\Sigma y = 280$, $\Sigma x^2 = 32,393$, $\Sigma y^2 = 13,142$, $\Sigma xy = 20,599$, and $r \approx 0.784$.
- (b) Use a 5% level of significance to test the claim that $\rho > 0$.
- (c) Verify that $S_e \approx 2.6964$, $a \approx 16.542$, $b \approx 0.4117$, and $\bar{x} \approx 73.167$.
- (d) Find the predicted percentage \hat{y} of successful field goals for a player with $x = 70\%$ successful free throws.
- (e) Find a 90% confidence interval for y when $x = 70$.
- (f) Use a 5% level of significance to test the claim that $\beta > 0$.
- (g) Find a 90% confidence interval for β and *interpret* its meaning.
8. **Baseball: Batting Average and Strikeouts** Let x be a random variable that represents the batting average of a professional baseball player. Let y be a random variable that represents the percentage of strikeouts of a professional baseball player. A random sample of $n = 6$ professional baseball players gave the following information (Reference: *The Baseball Encyclopedia*, Macmillan).

x	0.328	0.290	0.340	0.248	0.367	0.269
y	3.2	7.6	4.0	8.6	3.1	11.1

- (a) Verify that $\Sigma x = 1.842$, $\Sigma y = 37.6$, $\Sigma x^2 = 0.575838$, $\Sigma y^2 = 290.78$, $\Sigma xy = 10.87$, and $r \approx -0.891$.
- (b) Use a 5% level of significance to test the claim that $\rho \neq 0$.
- (c) Verify that $S_e \approx 1.6838$, $a \approx 26.247$, and $b \approx -65.081$.
- (d) Find the predicted percentage of strikeouts for a player with an $x = 0.300$ batting average.
- (e) Find an 80% confidence interval for y when $x = 0.300$.
- (f) Use a 5% level of significance to test the claim that $\beta \neq 0$.
- (g) Find a 90% confidence interval for β and *interpret* its meaning.
9. **Scuba Diving: Depth** What is the optimal amount of time for a scuba diver to be on the bottom of the ocean? That depends on the depth of the dive. The U.S. Navy has done a lot of research on this topic. The Navy defines the “optimal time” to be the time at each depth for the best balance between length of work period and decompression time after surfacing. Let x = depth of dive in meters, and let y = optimal time in hours. A random sample of divers gave the following data (based on information taken from *Medical Physiology* by A. C. Guyton, M.D.).

x	14.1	24.3	30.2	38.3	51.3	20.5	22.7
y	2.58	2.08	1.58	1.03	0.75	2.38	2.20

- (a) Verify that $\Sigma x = 201.4$, $\Sigma y = 12.6$, $\Sigma x^2 = 6734.46$, $\Sigma y^2 = 25.607$, $\Sigma xy = 311.292$, and $r \approx -0.976$.
- (b) Use a 1% level of significance to test the claim that $\rho < 0$.
- (c) Verify that $S_e \approx 0.1660$, $a \approx 3.366$, and $b \approx -0.0544$.
- (d) Find the predicted optimal time in hours for a dive depth of $x = 18$ meters.
- (e) Find an 80% confidence interval for y when $x = 18$ meters.
- (f) Use a 1% level of significance to test the claim that $\beta < 0$.
- (g) Find a 90% confidence interval for β and *interpret* its meaning.
10. **Physiology: Oxygen** Aviation and high-altitude physiology is a specialty in the study of medicine. Let x = partial pressure of oxygen in the alveoli (air cells in the lungs) when breathing naturally available air. Let y = partial pressure when breathing pure oxygen. The (x, y) data pairs correspond to elevations from 10,000 feet to 30,000 feet in 5000-foot intervals for a random sample of volunteers. Although the medical data were collected using airplanes, they apply equally well to Mt. Everest climbers (summit 29,028 feet).

x	6.7	5.1	4.2	3.3	2.1 (units: mm Hg/10)
y	43.6	32.9	26.2	6.2	13.9 (units: mm Hg/10)

- (Based on information taken from *Medical Physiology* by A. C. Guyton, M.D.)
- (a) Verify that $\Sigma x = 21.4$, $\Sigma y = 132.8$, $\Sigma x^2 = 103.84$, $\Sigma y^2 = 4125.46$, $\Sigma xy = 652.6$, and $r \approx 0.984$.
- (b) Use a 1% level of significance to test the claim that $\rho > 0$.
- (c) Verify that $S_e \approx 2.5319$, $a \approx -2.869$, and $b \approx 6.876$.
- (d) Find the predicted pressure when breathing pure oxygen if the pressure from breathing available air is $x = 4.0$.
- (e) Find a 90% confidence interval for y when $x = 4.0$.
- (f) Use a 1% level of significance to test the claim that $\beta > 0$.
- (g) Find a 95% confidence interval for β and *interpret* its meaning.
11. **New Car: Negotiating Price** Suppose you are interested in buying a new Toyota Corolla. You are standing on the sales lot looking at a model with different options. The list price is on the vehicle. As a salesperson approaches, you wonder what the dealer invoice price is for this model with its options. The following data are based on information taken from *Consumer Guide* (Vol. 677). Let x be the list price (in thousands of dollars) for a random selection of Toyota Corollas of different models and options. Let y be the dealer invoice (in thousands of dollars) for the given vehicle.

x	12.6	13.0	12.8	13.6	13.4	14.2
y	11.6	12.0	11.5	12.2	12.0	12.8

- (a) Verify that $\Sigma x = 79.6$, $\Sigma y = 72.1$, $\Sigma x^2 = 1057.76$, $\Sigma y^2 = 867.49$, $\Sigma xy = 957.84$, and $r \approx 0.956$.
- (b) Use a 1% level of significance to test the claim that $\rho > 0$.
- (c) Verify that $S_e \approx 0.1527$, $a \approx 1.965$, and $b \approx 0.758$.
- (d) Find the predicted dealer invoice when the list price is $x = 14$ (thousand dollars).
- (e) Find an 85% confidence interval for y when $x = 14$ (thousand dollars).
- (f) Use a 1% level of significance to test the claim that $\beta > 0$.
- (g) Find a 95% confidence interval for β and *interpret* its meaning.

12. **New Car: Negotiating Price** Suppose you are interested in buying a new Lincoln Navigator or Town Car. You are standing on the sales lot looking at a model with different options. The list price is on the vehicle. As a salesperson approaches, you wonder what the dealer invoice price is for this model with its options. The following data are based on information taken from *Consumer Guide* (Vol. 677). Let x be the list price (in thousands of dollars) for a random selection of these cars of different models and options. Let y be the dealer invoice (in thousands of dollars) for the given vehicle.

x	32.1	33.5	36.1	44.0	47.8
y	29.8	31.1	32.0	42.1	42.2

- (a) Verify that $\Sigma x = 193.5$, $\Sigma y = 177.2$, $\Sigma x^2 = 7676.71$, $\Sigma y^2 = 6432.5$, $\Sigma xy = 7023.19$, and $r \approx 0.977$.
- (b) Use a 1% level of significance to test the claim that $\rho > 0$.
- (c) Verify that $S_e \approx 1.5223$, $a \approx 1.4084$, and $b \approx 0.8794$.
- (d) Find the predicted dealer invoice when the list price is $x = 40$ (thousand dollars).
- (e) Find a 95% confidence interval for y when $x = 40$ (thousand dollars).
- (f) Use a 1% level of significance to test the claim that $\beta > 0$.
- (g) Find a 90% confidence interval for β and *interpret* its meaning.



13. **Expand Your Knowledge: Sample Size and Significance of r**

- (a) Suppose $n = 6$ and the sample correlation coefficient is $r = 0.90$. Is r significant at the 1% level of significance (based on a two-tailed test)?
- (b) Suppose $n = 10$ and the sample correlation coefficient is $r = 0.90$. Is r significant at the 1% level of significance (based on a two-tailed test)?
- (c) Explain why the test results of parts (a) and (b) are different even though the sample correlation coefficient $r = 0.90$ is the same in both parts. Does it appear that sample size plays an important role in determining the significance of a correlation coefficient? Explain.



14. **Expand Your Knowledge: Student's t Value for Sample r and for Sample b** It is not obvious from the formulas, but the values of the sample test statistic t for the correlation coefficient and for the slope of the least-squares line are equal for the same data set. This fact is based on the relation

$$b = r \frac{s_y}{s_x}$$

where s_y and s_x are the sample standard deviations of the x and y values, respectively.

- (a) Many computer software packages give the t value and corresponding P -value for b . If β is significant, is ρ significant?
- (b) When doing statistical tests “by hand,” it is easier to compute the sample test statistic t for the sample correlation coefficient r than it is to compute the sample test statistic t for the slope b of the sample least-squares line. Compare the results of parts (b) and (f) for Problems 7–12 of this problem set. Is the sample test statistic t for r the same as the corresponding test statistic for b ? If you conclude that ρ is positive, can you conclude that β is positive at the same level of significance? If you conclude that ρ is not significant, is β also not significant at the same level of significance?

SECTION 9.4

Multiple Regression

FOCUS POINTS

- Learn about the advantages of multiple regression.
- Learn the basic ingredients that go into a multiple regression model.
- Discuss standard error for computed coefficients and the coefficient of multiple determination.
- Test coefficients in the model for statistical significance.
- Compute confidence intervals for predictions.

Advantages of Multiple Regression

There are many examples in statistics in which one variable can be predicted very accurately in terms of another *single* variable. However, predictions usually improve if we consider additional relevant information. For example, the sugar content y of golden delicious apples taken from an apple orchard in Colorado could be predicted from $x_1 =$ number of days in growing season. If we also included information regarding $x_2 =$ soil quality rating and $x_3 =$ amount of available water, then we would expect our prediction of $y =$ sugar content to be more accurate.

Likewise, the annual net income y of a new franchise auto parts store could be predicted using only $x_1 =$ population size of sales district. However, we would probably get a better prediction of y values if we included the explanatory variables $x_2 =$ size of store inventory, $x_3 =$ dollar amount spent on advertising in local newspapers, and $x_4 =$ number of competing stores in the sales district.

For most statistical applications, we gain a definite advantage in the reliability of our predictions if we include more *relevant* data and corresponding (relevant) random variables in the computation of our predictions. In this section, we will give you an idea of how this can be done by methods of *multiple regression*. You should be aware that an in-depth study of multiple regression requires the use of advanced mathematics. However, if you are willing to let the computer be a “friend who gives you useful information,” then you will learn a great deal about multiple regression in this section. We will let the computer do most of the calculating work while we interpret the results.

Multiple regression

Basic Terminology and Notation

In statistics, the most commonly used mathematical formulas for expressing linear relationships among more than two variables are *equations* of the form

$$y = b_0 + b_1x_1 + b_2x_2 + \cdots + b_kx_k \quad (9)$$

Here, y is the variable that we want to predict or forecast. We will employ the usual terminology and call y the *response variable*. The k variables x_1, x_2, \dots, x_k are specified variables on which the predictions are going to be based. Once again, we will employ the popular terminology and call x_1, x_2, \dots, x_k the *explanatory variables*. This terminology is easy to remember if you just think of the explanatory variables x_1, x_2, \dots, x_k as “explaining” the response y .

In Equation (9), $b_0, b_1, b_2, \dots, b_k$ are numerical constants (called *coefficients*) that must be mathematically determined from given data. The numerical values of these coefficients are obtained from the *least-squares criterion*, which we will discuss after the following exercise.

GUIDED EXERCISE 9

Components of multiple regression equation

An industrial psychologist working for a hospital-supply company is studying the following variables for a random sample of company employees:

x_1 = number of years the employee has been with the company

x_2 = job-training level (0 = lowest level and 5 = highest level)

x_3 = interpersonal skills (0 = lowest level and 10 = highest level)

y = job-performance rating from supervisor (1 = lowest rating, 20 = highest rating)

The psychologist wants to predict y using x_1 , x_2 , and x_3 together in a least-squares equation.

- (a) Identify the response variable and the explanatory variables.



The response variable is what we want to predict. This is y , job performance. The explanatory variables are years of experience x_1 , training level x_2 , and interpersonal skills x_3 . In a sense, these variables “explain” the response variable.

- (b) After collecting data, the psychologist used a computer with appropriate software to obtain the least-squares linear equation

$$y = 1 + 0.2x_1 + 2.3x_2 + 0.7x_3$$

Identify the constant term and each of the coefficients with its corresponding variable.



The constant term is 1.

Explanatory Variable	Coefficient
x_1	0.2
x_2	2.3
x_3	0.7

- (c) Use the equation to predict the job-performance rating of an employee with 3 years of experience, a training level of 4, and an interpersonal skill rating of 2.



Substituting $x_1 = 3$, $x_2 = 4$, and $x_3 = 2$ into the least-squares equation and multiplying by the respective coefficients, we obtain the predicted job performance rating of

$$y = 1 + 0.2(3) + 2.3(4) + 0.7(2) = 12.2$$

Of course, the *predicted* value for job performance might differ from the actual rating given by the supervisor.

Theory for the least-squares criterion (optional)

This material is a little sophisticated, so you may wish to skip ahead to the discussion of regression models and computers and omit the following explanation of basic theory.

In multiple regression, the least-squares criterion states that the following sum (over all data points),

$$\sum [y_i - (b_0 + b_1x_{1i} + b_2x_{2i} + \cdots + b_kx_{ki})]^2 \quad (10)$$

must be made as small as possible. In this formula,

y_i = i th data value for y

x_{1i} = i th data value for x_1

x_{2i} = i th data value for x_2

\vdots

x_{ki} = i th data value for x_k

Recall that Equation (9) gives the predicted y value; therefore,

$$y_i - (b_0 + b_1x_{1i} + b_2x_{2i} + \cdots + b_kx_{ki}) \quad (11)$$

Residual

represents the *difference* between the *observed* y value (that is, y_i) and the *predicted* y value based on the data values $x_{1i}, x_{2i}, \dots, x_{ki}$. When we square this difference, total the result over all data points, and choose the values of $b_0, b_1, b_2, \dots, b_k$ to minimize the sum [i.e., minimize Equation (10)], then we are satisfying the least-squares criterion.

COMMENT The algebraic expression in Equation (11) is very important. In fact, it has a special name in the theory of regression. It is called a *residual*. The residual is simply the difference between the actual data value and the predicted value of the response variable based on given data values for the explanatory variables. Advanced topics in the theory of regression study residuals in great detail. Such a detailed treatment is beyond the scope of this text. However, from the discussion presented so far, we see that the method of least squares chooses the values of the coefficients b_i to make the sum of the squares of the residuals as small as possible.

After a good deal of mathematics has been done (involving a considerable amount of calculus), the least-squares criterion can be reduced to solving a system of linear equations. These are usually called *normal equations* (not to be confused with the normal distribution).

In the simplest case, where there are only *two* explanatory variables x_1 and x_2 and we want to fit the equation

$$y = b_0 + b_1x_1 + b_2x_2$$

to given data, there are three normal equations that must be solved for $b_0, b_1,$ and b_2 . These normal equations are

$$\left. \begin{aligned} \Sigma y_i &= nb_0 + b_1(\Sigma x_{1i}) + b_2(\Sigma x_{2i}) \\ \Sigma x_{1i}y_i &= b_0(\Sigma x_{1i}) + b_1(\Sigma x_{1i}^2) + b_2(\Sigma x_{1i}x_{2i}) \\ \Sigma x_{2i}y_i &= b_0(\Sigma x_{2i}) + b_1(\Sigma x_{1i}x_{2i}) + b_2(\Sigma x_{2i}^2) \end{aligned} \right\} \quad (12)$$

In the system of Equations (12), n represents the number of data points and $x_{1i}, x_{2i},$ and y_i all represent given data values.

Therefore, the only unknowns are the coefficients $b_0, b_1,$ and b_2 ; we can use the system of Equations (12) to solve for these unknowns. This is the procedure that lets us obtain the least-squares regression equation in Equation (9) when we have only *two* explanatory variables.

As you can see, this is all rather complicated, and the more explanatory variables x_1, x_2, \dots, x_k we have, the more involved the calculations become. In the general case, if you have k explanatory variables, there will be $k + 1$ normal equations that must be solved for the coefficients $b_0, b_1, b_2, \dots, b_k$.

Regression Models and Computers

As you can see from the preceding optional discussion, the work required to find an equation satisfying the least-squares criterion is tremendous and can be very complex. Today, such work is conveniently left to computers. In this text, we use two computer software packages that specialize in statistical applications.

Minitab is a widely used statistical software package. It fully supports multiple regression. Excel 2007 has a multiple regression component that performs much of the multiple regression analysis. We will use Minitab in our example. Many other software packages, including SPSS, support multiple regression and have outputs similar to those of Minitab.

Ingredients of the regression model

In this section, we will often refer to a *regression model*. What do we mean by this? We mean a mathematical package that consists of the following ingredients:

1. The model will have a collection of random variables, *one* of which has been identified as the response variable, with *any or all* of the remaining variables being identified as explanatory variables.
2. Associated with a given application will be a collection of numerical data values for each of the variables of part 1.
3. Using the numerical data values, the least-squares criterion, and the declared response and explanatory variables, a *least-squares equation* (also called a *regression equation*) will be constructed. In Section 9.2, we were able to construct the least-squares equation using only a hand calculator. However, in multiple regression, we will use a computer to construct the least-squares equation.
4. The model usually includes additional information about the variables used, the coefficients and regression equation, and a measure of “goodness of fit” of the regression equation to the data values. In modern practice, this information usually comes to you in the form of computer displays.
5. Finally, the regression model enables you to supply given values of the explanatory variables for the purpose of predicting or forecasting the corresponding value of the response variable. You also should be able to construct a *c%* confidence interval for your least-squares prediction. In multiple regression, this will be done by the computer at your request.



Problem 7 at the end of this section discusses *curvilinear regression* (also known as *polynomial regression*).

The next example demonstrates computer applications of a typical multiple regression problem. In the context of the example, we will introduce some of the basic techniques of multiple regression.

Example Utilizing Minitab

EXAMPLE 10

MULTIPLE REGRESSION

Antelope are beautiful and graceful animals that live on the high plains of the western United States. Thunder Basin National Grasslands in Wyoming is home to hundreds of antelope. The Bureau of Land Management (BLM) has been studying the Thunder Basin antelope population for the past 8 years. The variables used are

- x_1 = spring fawn count (in hundreds of fawns)
- x_2 = size of adult antelope population (in hundreds)
- x_3 = annual precipitation (in inches)
- x_4 = winter severity index (1 = mild and 5 = extremely severe) (This is an index based on temperature and wind chill factors.)

The data obtained in the study over the 8-year period are shown in Table 9-15.

TABLE 9-15 Data for Thunder Basin Antelope Study

Year	x_1	x_2	x_3	x_4
1	2.9	9.2	13.2	2
2	2.4	8.7	11.5	3
3	2.0	7.2	10.8	4
4	2.3	8.5	12.3	2
5	3.2	9.6	12.6	3
6	1.9	6.8	10.6	5
7	3.4	9.7	14.1	1
8	2.1	7.9	11.2	3



D. Robert & Lorri Franz/CORBIS

Summary Statistics for Each Variable

It is a good idea to first look at the summary statistics for each variable. Figure 9-19 shows the Minitab display of the summary statistics.

Menu selection: Stat ► Basic Statistic ► Display Descriptive Statistics

FIGURE 9-19

Minitab Display of Summary Statistics for Each Variable

Descriptive Statistics						
Variable	N	Mean	Median	TrMean	StDev	SE Mean
x1	8	2.525	2.350	2.525	0.570	0.202
x2	8	8.450	8.600	8.450	1.076	0.380
x3	8	12.037	11.900	12.037	1.229	0.435
x4	8	2.875	3.000	2.875	1.246	0.441
Variable	Minimum	Maximum	Q1	Q3		
x1	1.900	3.400	2.025	3.125		
x2	6.800	9.700	7.375	9.500		
x3	10.600	14.100	10.900	13.050		
x4	1.000	5.000	2.000	3.750		

This type of information can be very useful because it tells you basic information about the variables you are studying. Sample means and sample standard deviations with a Student's t distribution are essential ingredients for estimating or testing population means (Chapters 7 and 8).

For example, if μ_2 represents the *population mean* of x_2 (adult antelope population), then by using the methods of Section 7.2 we can quickly estimate a 90% confidence interval for μ_2 :

$$7.729 < \mu_2 < 9.171$$

Since our units are in hundreds, this means we can be 90% sure that the *population mean* μ_2 of adult antelope in the Thunder Basin National Grasslands is between 773 and 917.

Correlation Between Variables

It is also useful to examine how the variables relate to each other. Figure 9-20 shows the sample correlation coefficients r between each of the two variables. A natural question arises: Which of the variables are closely related to each other, and which are not as closely related? Recall (from Section 9.1) that if the correlation coefficient is near 1 or -1 , then the corresponding variables have a lot in common. If the correlation coefficient is near zero, the variables have much less influence on each other.

Menu selection: Stat ► Basic Statistics ► Correlation

FIGURE 9-20

Minitab Display of Correlation Coefficients Between Variables

Correlations (Pearson)			
	x1	x2	x3
x2	0.939		
x3	0.924	0.903	
x4	-0.739	-0.836	-0.901

Look at Figure 9-20. Which of the variables has the greatest influence on x_1 ? The sample correlation coefficient between x_1 and x_2 is $r = 0.939$, with a corresponding coefficient of determination of $r^2 \approx 0.88$. This means that if we consider only x_1 and x_2 (and none of the other variables), then about 88% of the variation in x_1 can be explained by the corresponding variation in x_2 (by itself). Similarly, if we consider only x_1 and x_3 , we see the sample correlation coefficient $r = 0.924$, with a corresponding coefficient of determination of $r^2 \approx 0.85$. About

85% of the variation in x_1 can be explained by the corresponding variation in x_3 . The variable x_4 has much less influence on x_1 because the sample correlation coefficient between these two variables is $r = -0.739$, with corresponding coefficient of determination $r^2 \approx 0.55$, or only 55%.

These relationships are very reasonable in the context of our problem. It is common sense that the number of spring fawns x_1 is strongly related to x_2 , the size of the adult antelope population. Furthermore, the spring fawn count x_1 is very much influenced by available food for the fawn (and its mother). Thunder Basin National Grasslands is a semiarid region, and available food (grass) is almost completely determined by annual precipitation x_3 . Antelope are naturally strong and hardy animals. Therefore, the temperature and wind chill index x_4 will have much less effect on the adult does and corresponding number of spring fawns provided there is plenty of available food.

Least-Squares Equation

Figure 9-21 shows a display that gives an expression for the actual least-squares equation and a lot of information about the equation. To get this display or a similar display, the user needs to declare which variable is the response variable and which are the explanatory variables. For Figure 9-21, we designated x_1 as the response variable. This means that x_1 is the variable we choose to predict. We also designated variables x_2 , x_3 , and x_4 as explanatory variables. This means that x_2 , x_3 , and x_4 will be used *together* to predict x_1 . There is a lot of flexibility here. We could have designated any *one* of the variables x_1 , x_2 , x_3 , x_4 as the response variable and *any or all* of the remaining variables as explanatory variables. So there are several possible regression models the computer can construct for you, depending on the type of information you want. In this example, we want to predict x_1 (spring fawn count) by using x_2 (adult population), x_3 (annual precipitation), and x_4 (winter index) *together*.

Menu selection: **Stat** ► **Regression** ► **Regression**. In the dialogue box, select x_1 as the response and x_2 , x_3 , x_4 as the predictors.

FIGURE 9-21

Minitab Display of Regression Analysis

Regression Analysis				
The regression equation is				
$x_1 = -5.92 + 0.338 x_2 + 0.402 x_3 + 0.263 x_4$				
Predictor	Coef	StDev	T	P
Constant	-5.922	1.256	-4.72	0.009
x2	0.33822	0.09947	3.40	0.027
x3	0.4015	0.1099	3.65	0.022
x4	0.26295	0.08514	3.09	0.037
S = 0.1209	R-Sq = 97.4%	R-Sq(adj) = 95.5%		

The least-squares regression equation is given near the top of the display. Then more information is given about the constant and coefficients. The parts of the equation are

$$\begin{array}{cccccc}
 x_1 = & -5.92 & + & 0.338x_2 & + & 0.402x_3 & + & 0.263x_4 & & (13) \\
 \uparrow & & \uparrow & \swarrow & \uparrow & \nearrow & & & & \\
 \text{response} & & \text{constant} & & \text{coefficient of} & & & & & \\
 \text{variable} & & & & \text{associated explanatory} & & & & & \\
 & & & & \text{variable} & & & & &
 \end{array}$$

COMMENT In the case of a simple regression model in which we have only one explanatory variable, the coefficient of that variable is the *slope* of the least-squares line. This slope (or coefficient) represents the change in the response variable per unit change in the explanatory variable. In a multiple regression model such as Equation (13), the coefficients also can be thought of as a slope, *provided* we hold the other variables as arbitrary and fixed constants. For

example, the coefficient of x_2 in Equation (13) is $b_2 = 0.338$. This means that if x_3 (precipitation) and x_4 (winter index) are taken into account but held constant, then $b_2 = 0.338$ represents the change in x_1 (spring fawn count) per unit change in x_2 (adult antelope count). Since our units are in hundreds, this indicates that if x_3 and x_4 are taken into account as arbitrary but fixed values, then an increase of 100 adult antelope would give an expected increase of 33.8, or 34, spring fawns.

A natural question arises: How good a fit is the least-squares regression Equation (13) for our given data?

Coefficient of multiple determination

One way to answer this question is to examine the *coefficient of multiple determination*. The coefficient of multiple determination is a direct generalization of the concept of coefficient of determination (between *two* variables) as discussed in Section 9.2, and it has essentially the same meaning. The coefficient of multiple determination is given in the display of Figure 9-21 as a percent. We see $R\text{-sq} = 97.4\%$. This means that about 97.4% of the variation in the response variable x_1 can be explained from the least-squares Equation (13) and the corresponding *joint* variation of the variables x_2 , x_3 , and x_4 taken together. The remaining $100\% - 97.4\% = 2.6\%$ of the variation in x_1 is due to random chance or possibly the presence of other variables not included in this regression equation. (We will discuss the *standard error* associated with each coefficient later in this section.)

Predictions

Let's use the current regression model to predict the response variable x_1 . Recall that in Section 9.2 we first made predictions from the least-squares line and then constructed a confidence interval for our predictions. Although the exact details are beyond the scope of this text, this process can be generalized to multiple regression. The calculations are very tedious, but that's why we use a computer!

Suppose we ask the following question: In a year when $x_2 = 8.2$ (hundreds of adult antelope), $x_3 = 11.7$ (inches of precipitation), and $x_4 = 3$ (winter index), what do we predict for x_1 (spring fawn count)? Furthermore, let's suppose we want an 85% confidence interval for our prediction.

To answer this question, we look at Figure 9-22, which shows the Minitab prediction result for x_1 from the specified values of x_2 , x_3 , and x_4 .

Menu selection: **Stat** ► **Regression** ► **Regression**. In the dialogue box, select Options. List the new observations for x_2 , x_3 , and x_4 in order, separated by spaces. Specify the confidence level. Be sure that Fit Intercept is checked.

FIGURE 9-22

Minitab Display Showing the Predicted Value of x_1

Predicted Values				
Fit	StDev	Fit	85.0% CI	85.0% PI
2.3378	0.0472		(2.2539, 2.4217)	(2.1069, 2.5687)

The value for Fit is 2.3378. This is the predicted value for x_1 . The 85% confidence interval for the prediction is designated as 85% PI. We see that the interval for x_1 (rounded to two digits after the decimal) is $2.11 \leq x_1 \leq 2.57$. This means we are 85% confident that the number of spring fawns will be in the range of 211 to 257.

Please note that this is *not* a confidence interval for the population mean of x_1 . Rather, we have constructed a confidence interval for the *actual value* of x_1 under the conditions $x_2 = 8.2$, $x_3 = 11.7$, and $x_4 = 3$.

COMMENT Extrapolation much beyond the data range for any of the variables in a multiple regression model can produce results that might be meaningless and unrealistic. Many computer software packages warn about computing a confidence interval for a prediction when some of the values of the explanatory variables are beyond the data range in either direction.

Testing a Coefficient for Significance

In applications of multiple regression, it is possible to have many different variables. Occasionally, you might suspect that one of the explanatory variables x_i is not very useful as a tool for predicting the response variable. It simply may not influence the response variable much at all. To decide whether or not this is the case, we construct a test for the significance of the coefficient of x_i in the least-squares equation.

Recall that the general least-squares equation is

$$y = b_0 + b_1x_1 + b_2x_2 + \cdots + b_kx_k \quad (14)$$

where y = response variable

x_i = explanatory variable for $i = 1, 2, \dots, k$

b_i = numerical coefficient for $i = 0, 1, 2, \dots, k$

Equation (14) was constructed from given data. Usually, the data are only a small subset of all possible data that could have been collected.

Let us suppose (in theory) that we used *all possible data* that could ever be obtained for our regression problem and that we constructed the regression equation using the entire population of all possible data. Then we would get the *theoretical* regression equation

$$y = \beta_0 + \beta_1x_1 + \beta_2x_2 + \cdots + \beta_kx_k \quad (15)$$

where y and x_i are as in Equation (14), but β_i is the *theoretical* coefficient of x_i .

Now look back at the regression analysis in Figure 9-21. Beside the constant and each coefficient is a number in the StDev column. This is the *standard error* corresponding to that coefficient. The standard error can be thought of as similar to a standard deviation that corresponds to the coefficient. The calculation of the number is beyond the scope of this text, but it is available on computer printouts, and we will use it to construct our test.

Let us call S_i the standard error for coefficient x_i (S_0 is the standard error for the constant). Under very basic and general assumptions, it can be proved that

$$t = \frac{b_i - \beta_i}{S_i} \quad (16)$$

has a Student's t distribution with degrees of freedom $d.f. = n - k - 1$, where n = number of data points and k = number of explanatory variables in the least-squares equation.

Now let us return to the question: Is x_i useful as an explanatory variable in the least-squares equation?

The answer is that it is *not* useful if $\beta_i = 0$. In that case, the (theoretical) coefficient of x_i would be zero and x_i would contribute nothing to the least-squares equation. However, if $\beta_i \neq 0$, then the explanatory variable x_i does contribute information in the least-squares equation.

Consider the following hypotheses,

$$H_0: \beta_i = 0 \quad \text{and} \quad H_1: \beta_i \neq 0$$

If we accept H_0 , we conclude that $\beta_i = 0$ and x_i probably should be dropped as an explanatory variable in the least-squares equation. If we accept H_1 , we conclude that $\beta_i \neq 0$ and x_i should be included as an explanatory variable in the least-squares equation.

EXAMPLE 11 TESTING A COEFFICIENT

We'll use the data and printouts of Example 10 and test the significance of x_3 as an explanatory variable using level of significance $\alpha = 0.05$.

$$H_0: \beta_3 = 0 \quad \text{and} \quad H_1: \beta_3 \neq 0$$

To find the t value corresponding to b_3 , we use Equation (16) and the null hypothesis $H_0: \beta_3 = 0$. This gives us the equation

$$t = \frac{b_3}{S_3} \quad (17)$$

In the regression analysis shown in Figure 9-21, we see a t value for the constant and each coefficient. This t value is exactly the value of $t = b_i/S_i$. This is the t value corresponding to the sample test statistic. For the coefficient of x_3 , we see

$$t \text{ value} \approx 3.65$$

Notice in Figure 9-21 that we are also given the P -value based on a two-tailed test of the sample test statistic for each coefficient. This is the value in the column headed "P." For the sample test statistic $t \approx 3.65$, the corresponding P -value is 0.022. Since the P -value is less than the level of significance $\alpha = 0.05$, we reject H_0 . In other words, at the 5% level of significance, we can say that the population correlation coefficient β_3 of x_3 is not 0.

We conclude at the 5% level of significance that x_3 (annual precipitation) should be included as an explanatory variable in the least-squares equation. Notice that Figure 9-21 also gives the P -value for each ratio, so we can conclude the test using P -values directly. Using the P -values, we see that x_2 and x_3 are also significant at the 5% level.

Confidence Intervals for Coefficients

Equation (16) also gives us the basis for finding *confidence intervals* for β_i . A $c\%$ confidence interval for β_i will be

$$b_i - tS_i < \beta_i < b_i + tS_i$$

where $d.f. = n - k - 1$, t is selected according to the specified confidence level, b_i is the numerical value of the coefficient from Figure 9-21, S_i is the numerical value of the standard error from Figure 9-21, n is the number of data points, and k is the number of explanatory variables in the least-squares equation.

EXAMPLE 12 CONFIDENCE INTERVAL FOR A COEFFICIENT

Suppose we want to compute a 90% confidence interval for β_2 , the coefficient of x_2 . From Figure 9-21, we have (rounding to three digits after the decimal)

$$b_2 = 0.338, \quad S_2 = 0.099, \quad \text{and} \quad d.f. = 4$$

From the t table (Table 6, Appendix II), we find $t = 2.132$, so,

$$\begin{aligned} b_2 - tS_2 &< \beta_2 < b_2 + tS_2 \\ 0.338 - 2.132(0.099) &< \beta_2 < 0.338 + 2.132(0.099) \\ 0.127 &< \beta_2 < 0.549 \end{aligned}$$

Excel 2007 Displays

Although Excel gives information very similar to that supplied by Minitab, the least-squares equation is not explicitly displayed. However, the intercept (constant) and coefficients of the variables are shown with the corresponding standard errors and t values with P -values (two-tailed test). Excel shows the confidence interval for each coefficient. However, there is no built-in function to provide predicted values or confidence intervals for predicted values. Note that as in the Minitab regression analysis, we will not make use of the ANOVA information in the Excel display.

On the home screen, click the **Data** tab. Select **Data Analysis** from the Analysis group. In the dialogue box, select **Regression**. Note that when you enter data into the worksheet, all the explanatory variables must be together in a block. Figure 9-23 shows the Excel display for Examples 10 and 11.

FIGURE 9-23

Excel Display of Regression Analysis

Regression Statistics						
Multiple R	0.987060478					
R Square	0.974288388					
Adjusted R Square	0.955004679					
Standard Error	0.120927579					
Observations	8					
ANOVA						
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>	
Regression	3	2.216506083	0.738835361	50.5239104	0.001228863	
Residual	4	0.058493917	0.014623479			
Total	7	2.275				
	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	-5.922011616	1.255623292	-4.716391972	0.009196085	-9.40818798	-2.435835251
x2	0.338217487	0.099470083	3.400193085	0.027272474	0.062043691	0.614391283
x3	0.401503945	0.109900277	3.653347874	0.021707246	0.096371226	0.706636664
x4	0.262946128	0.085136028	3.088541172	0.036626194	0.02657013	0.499322125

VIEWPOINT

Synoptic Climatology

Synoptic means “giving a summary from the same basic point of view.” In this case, the point of view is Nivot Ridge, high above the timberline in the Rocky Mountains. Vegetation, water, temperature, and wind all affect the delicate balance of this alpine environment. How do these elements of nature interact to sustain life in such a harsh land? One answer can be found by collecting data at the location and using multiple regression to study the interaction of variables. For more information, visit the Brase/Brase statistics site at <http://www.cengage.com/statistics/brase> and find the link to Nivot Ridge Climate Study.

SECTION 9.4 PROBLEMS

1. **Statistical Literacy** Given the linear regression equation

$$x_1 = 1.6 + 3.5x_2 - 7.9x_3 + 2.0x_4$$

- (a) Which variable is the response variable? Which variables are the explanatory variables?

- (b) Which number is the constant term? List the coefficients with their corresponding explanatory variables.
- (c) If $x_2 = 2$, $x_3 = 1$, and $x_4 = 5$, what is the predicted value for x_1 ?
- (d) Explain how each coefficient can be thought of as a “slope” under certain conditions. Suppose x_3 and x_4 were held at fixed but arbitrary values and x_2 was increased by 1 unit. What would be the corresponding change in x_1 ? Suppose x_2 increased by 2 units. What would be the expected change in x_1 ? Suppose x_2 decreased by 4 units. What would be the expected change in x_1 ?
- (e) Suppose that $n = 12$ data points were used to construct the given regression equation and that the standard error for the coefficient of x_2 is 0.419. Construct a 90% confidence interval for the coefficient of x_2 .
- (f) Using the information of part (e) and level of significance 5%, test the claim that the coefficient of x_2 is different from zero. Explain how the conclusion of this test would affect the regression equation.
2. **Statistical Literacy** Given the linear regression equation

$$x_3 = -16.5 + 4.0x_1 + 9.2x_4 - 1.1x_7$$

- (a) Which variable is the response variable? Which variables are the explanatory variables?
- (b) Which number is the constant term? List the coefficients with their corresponding explanatory variables.
- (c) If $x_1 = 10$, $x_4 = -1$, and $x_7 = 2$, what is the predicted value for x_3 ?
- (d) Explain how each coefficient can be thought of as a “slope.” Suppose x_1 and x_7 were held as fixed but arbitrary values. If x_4 increased by 1 unit, what would we expect the corresponding change in x_3 to be? If x_4 increased by 3 units, what would be the corresponding expected change in x_3 ? If x_4 decreased by 2 units, what would we expect for the corresponding change in x_3 ?
- (e) Suppose that $n = 15$ data points were used to construct the given regression equation and that the standard error for the coefficient of x_4 is 0.921. Construct a 90% confidence interval for the coefficient of x_4 .
- (f) Using the information of part (e) and level of significance 1%, test the claim that the coefficient of x_4 is different from zero. Explain how the conclusion has a bearing on the regression equation.

For Problems 3–6, use appropriate multiple regression software of your choice and enter the data. Note that the data are also available for download at the Online Study Center in formats for Excel, Minitab portable files, SPSS files, and ASCII files.



3. **Medical: Blood Pressure** The systolic blood pressure of individuals is thought to be related to both age and weight. For a random sample of 11 men, the following data were obtained:

Systolic Blood Pressure	Age (years)	Weight (pounds)	Systolic Blood Pressure	Age (years)	Weight (pounds)
x_1	x_2	x_3	x_1	x_2	x_3
132	52	173	137	54	188
143	59	184	149	61	188
153	67	194	159	65	207
162	73	211	128	46	167
154	64	196	166	72	217
168	74	220			

- (a) Generate summary statistics, including the mean and standard deviation of each variable. Compute the coefficient of variation (see Section 3.2) for each variable. Relative to its mean, which variable has the greatest spread

- of data values? Which variable has the smallest spread of data values relative to its mean?
- For each pair of variables, generate the sample correlation coefficient r . Compute the corresponding coefficient of determination r^2 . Which variable (other than x_1) has the greatest influence (by itself) on x_1 ? Would you say that both variables x_2 and x_3 show a strong influence on x_1 ? Explain your answer. What percent of the variation in x_1 can be explained by the corresponding variation in x_2 ? Answer the same question for x_3 .
 - Perform a regression analysis with x_1 as the response variable. Use x_2 and x_3 as explanatory variables. Look at the coefficient of multiple determination. What percentage of the variation in x_1 can be explained by the corresponding variations in x_2 and x_3 *taken together*?
 - Look at the coefficients of the regression equation. Write out the regression equation. Explain how each coefficient can be thought of as a slope. If age were held fixed, but a person put on 10 pounds, what would you expect for the corresponding change in systolic blood pressure? If a person kept the same weight but got 10 years older, what would you expect for the corresponding change in systolic blood pressure?
 - Test each coefficient to determine if it is zero or not zero. Use level of significance 5%. Why would the outcome of each test help us determine whether or not a given variable should be used in the regression model?
 - Find a 90% confidence interval for each coefficient.
 - Suppose Michael is 68 years old and weighs 192 pounds. Predict his systolic blood pressure, and find a 90% confidence range for your prediction (if your software produces prediction intervals).



4. **Education: Exam Scores** Professor Gill has taught general psychology for many years. During the semester, she gives three multiple-choice exams, each worth 100 points. At the end of the course, Dr. Gill gives a comprehensive final worth 200 points. Let x_1 , x_2 , and x_3 represent a student's scores on exams 1, 2, and 3, respectively. Let x_4 represent the student's score on the final exam. Last semester Dr. Gill had 25 students in her class. The student exam scores are shown below.

x_1	x_2	x_3	x_4	x_1	x_2	x_3	x_4	x_1	x_2	x_3	x_4
73	80	75	152	79	70	88	164	81	90	93	183
93	88	93	185	69	70	73	141	88	92	86	177
89	91	90	180	70	65	74	141	78	83	77	159
96	98	100	196	93	95	91	184	82	86	90	177
73	66	70	142	79	80	73	152	86	82	89	175
53	46	55	101	70	73	78	148	78	83	85	175
69	74	77	149	93	89	96	192	76	83	71	149
47	56	60	115	78	75	68	147	96	93	95	192
87	79	90	175								

Since Professor Gill has not changed the course much from last semester to the present semester, the preceding data should be useful for constructing a regression model that describes this semester as well.

- Generate summary statistics, including the mean and standard deviation of each variable. Compute the coefficient of variation (see Section 3.2) for each variable. Relative to its mean, would you say that each exam had about the same spread of scores? Most professors do not wish to give an exam that is extremely easy or extremely hard. Would you say that all of the exams were about the same level of difficulty? (Consider both means and spread of test scores.)
- For each pair of variables, generate the sample correlation coefficient r . Compute the corresponding coefficient of determination r^2 . Of the three

- exams 1, 2, and 3, which do you think had the most influence on the final exam 4? Although one exam had more influence on the final exam, did the other two exams still have a lot of influence on the final? Explain each answer.
- Perform a regression analysis with x_4 as the response variable. Use x_1 , x_2 , and x_3 as explanatory variables. Look at the coefficient of multiple determination. What percentage of the variation in x_4 can be explained by the corresponding variations in x_1 , x_2 , and x_3 taken together?
 - Write out the regression equation. Explain how each coefficient can be thought of as a slope. If a student were to study “extra hard” for exam 3 and increase his or her score on that exam by 10 points, what corresponding change would you expect on the final exam? (Assume that exams 1 and 2 remain “fixed” in their scores.)
 - Test each coefficient in the regression equation to determine if it is zero or not zero. Use level of significance 5%. Why would the outcome of each hypothesis test help us decide whether or not a given variable should be used in the regression equation?
 - Find a 90% confidence interval for each coefficient.
 - This semester Susan has scores of 68, 72, and 75 on exams 1, 2, and 3, respectively. Make a prediction for Susan’s score on the final exam and find a 90% confidence interval for your prediction (if your software supports prediction intervals).



5. **Entertainment: Movies** A motion picture industry analyst is studying movies based on epic novels. The following data were obtained for 10 Hollywood movies made in the past five years. Each movie was based on an epic novel. For these data, x_1 = first-year box office receipts of the movie, x_2 = total production costs of the movie, x_3 = total promotional costs of the movie, and x_4 = total book sales prior to movie release. All units are in millions of dollars.

x_1	x_2	x_3	x_4	x_1	x_2	x_3	x_4
85.1	8.5	5.1	4.7	30.3	3.5	1.2	3.5
106.3	12.9	5.8	8.8	79.4	9.2	3.7	9.7
50.2	5.2	2.1	15.1	91.0	9.0	7.6	5.9
130.6	10.7	8.4	12.2	135.4	15.1	7.7	20.8
54.8	3.1	2.9	10.6	89.3	10.2	4.5	7.9

- Generate summary statistics, including the mean and standard deviation of each variable. Compute the coefficient of variation (see Section 3.2) for each variable. Relative to its mean, which variable has the largest spread of data values? Why would a variable with a large coefficient of variation be expected to change a lot relative to its average value? Although x_1 has the largest standard deviation, it has the smallest coefficient of variation. How does the mean of x_1 help explain this?
- For each pair of variables, generate the sample correlation coefficient r . Compute the corresponding coefficient of determination r^2 . Which of the three variables x_2 , x_3 , and x_4 has the *least* influence on box office receipts? What percent of the variation in box office receipts can be attributed to the corresponding variation in production costs?
- Perform a regression analysis with x_1 as the response variable. Use x_2 , x_3 , and x_4 as explanatory variables. Look at the coefficient of multiple determination. What percentage of the variation in x_1 can be explained by the corresponding variations in x_2 , x_3 , and x_4 taken together?
- Write out the regression equation. Explain how each coefficient can be thought of as a slope. If x_2 (production costs) and x_4 (book sales) were held fixed but x_3 (promotional costs) was increased by \$1 million, what would you expect for the corresponding change in x_1 (box office receipts)?

- (e) Test each coefficient in the regression equation to determine if it is zero or not zero. Use level of significance 5%. Explain why book sales x_4 probably are not contributing much information in the regression model to forecast box office receipts x_1 .
- (f) Find a 90% confidence interval for each coefficient.
- (g) Suppose a new movie (based on an epic novel) has just been released. Production costs were $x_2 = 11.4$ million; promotion costs were $x_3 = 4.7$ million; book sales were $x_4 = 8.1$ million. Make a prediction for $x_1 =$ first-year box office receipts and find an 85% confidence interval for your prediction (if your software supports prediction intervals).
- (h) Construct a new regression model with x_3 as the response variable and x_1 , x_2 , and x_4 as explanatory variables. Suppose Hollywood is planning a new epic movie with projected box office sales $x_1 = 100$ million and production costs $x_2 = 12$ million. The book on which the movie is based had sales of $x_4 = 9.2$ million. Forecast the dollar amount (in millions) that should be budgeted for promotion costs x_3 and find an 80% confidence interval for your prediction.



6. **Franchise Business: Market Analysis** All Greens is a franchise store that sells house plants and lawn and garden supplies. Although All Greens is a franchise, each store is owned and managed by private individuals. Some friends have asked you to go into business with them to open a new All Greens store in the suburbs of San Diego. The national franchise headquarters sent you the following information at your request. These data are about 27 All Greens stores in California. Each of the 27 stores has been doing very well, and you would like to use the information to help set up your own new store. The variables for which we have data are

- x_1 = annual net sales, in thousands of dollars
 x_2 = number of square feet of floor display in store, in thousands of square feet
 x_3 = value of store inventory, in thousands of dollars
 x_4 = amount spent on local advertising, in thousands of dollars
 x_5 = size of sales district, in thousands of families
 x_6 = number of competing or similar stores in sales district

A sales district was defined to be the region within a 5-mile radius of an All Greens store.

x_1	x_2	x_3	x_4	x_5	x_6	x_1	x_2	x_3	x_4	x_5	x_6
231	3	294	8.2	8.2	11	65	1.2	168	4.7	3.3	11
156	2.2	232	6.9	4.1	12	98	1.6	151	4.6	2.7	10
10	0.5	149	3	4.3	15	398	4.3	342	5.5	16.0	4
519	5.5	600	12	16.1	1	161	2.6	196	7.2	6.3	13
437	4.4	567	10.6	14.1	5	397	3.8	453	10.4	13.9	7
487	4.8	571	11.8	12.7	4	497	5.3	518	11.5	16.3	1
299	3.1	512	8.1	10.1	10	528	5.6	615	12.3	16.0	0
195	2.5	347	7.7	8.4	12	99	0.8	278	2.8	6.5	14
20	1.2	212	3.3	2.1	15	0.5	1.1	142	3.1	1.6	12
68	0.6	102	4.9	4.7	8	347	3.6	461	9.6	11.3	6
570	5.4	788	17.4	12.3	1	341	3.5	382	9.8	11.5	5
428	4.2	577	10.5	14.0	7	507	5.1	590	12.0	15.7	0
464	4.7	535	11.3	15.0	3	400	8.6	517	7.0	12.0	8
15	0.6	163	2.5	2.5	14						

- (a) Generate summary statistics, including the mean and standard deviation of each variable. Compute the coefficient of variation (see Section 3.2) for each variable. Relative to its mean, which variable has the largest spread of

- data values? Which variable has the least spread of data values relative to its mean?
- (b) For each pair of variables, generate the sample correlation coefficient r . For all pairs involving x_1 , compute the corresponding coefficient of determination r^2 . Which variable has the greatest influence on annual net sales? Which variable has the least influence on annual net sales?
- (c) Perform a regression analysis with x_1 as the response variable. Use $x_2, x_3, x_4, x_5,$ and x_6 as explanatory variables. Look at the coefficient of multiple determination. What percentage of the variation in x_1 can be explained by the corresponding variations in $x_2, x_3, x_4, x_5,$ and x_6 taken together?
- (d) Write out the regression equation. If two new competing stores moved into the sales district but the other explanatory variables did not change, what would you expect for the corresponding change in annual net sales? Explain your answer. If you increased the local advertising by a thousand dollars but the other explanatory variables did not change, what would you expect for the corresponding change in annual net sales? Explain.
- (e) Test each coefficient to determine if it is or is not zero. Use level of significance 5%.
- (f) Suppose you and your business associates rent a store, get a bank loan to start up your business, and do a little research on the size of your sales district and the number of competing stores in the district. If $x_2 = 2.8,$ $x_3 = 250, x_4 = 3.1, x_5 = 7.3,$ and $x_6 = 2,$ use a computer to forecast $x_1 =$ annual net sales and find an 80% confidence interval for your forecast (if your software produces prediction intervals).
- (g) Construct a new regression model with x_4 as the response variable and $x_1, x_2, x_3, x_5,$ and x_6 as explanatory variables. Suppose an All Greens store in Sonoma, California, wants to estimate a range of advertising costs appropriate to its store. If it spends too little on advertising, it will not reach enough customers. However, it does not want to overspend on advertising for this type and size of store. At this store, $x_1 = 163, x_2 = 2.4, x_3 = 188, x_5 = 6.6,$ and $x_6 = 10.$ Use these data to predict x_4 (advertising costs) and find an 80% confidence interval for your prediction. At the 80% confidence level, what range of advertising costs do you think is appropriate for this store?



7. **Expand Your Knowledge: Curvilinear Polynomial Regression** In this section we studied multiple linear regression. Our basic linear model has been

$$y = b_0 + b_1x_1 + b_2x_2 + \cdots + b_kx_k$$

Since all the variables x_1, x_2, \dots, x_k are of first degree, this is an example of linear regression. However, the same basic methods of linear regression can be used for *curvilinear regression* (also known as *polynomial regression*). The interested reader can find a great deal of information on this topic in the book *Applied Numerical Methods* by Carnahan, Luther, and Wilkes from page 573 on.

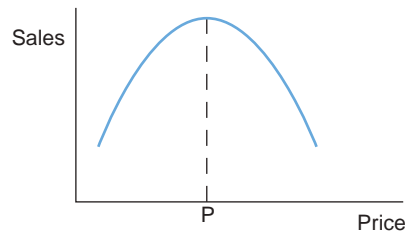
Assume we have at least $k + 2$ data pairs (x, y) and we want to approximate y using a polynomial of degree k . To do this, we make the following identification.

$$x_1 = x; x_2 = x^2; x_3 = x^3; \dots; x_k = x^k$$

Then we use our known methods of multiple regression to obtain coefficients $b_0, b_1, b_2, b_3, \dots, b_k$ and the equation $y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \cdots + b_kx_k$. This is called the *least-squares curvilinear regression model*.

Marketing studies show that price increases often have a point of diminishing returns. For a popular product, the price can often increase with sales.

However, when the price becomes too high, sales start to drop off. In the following graph, P = point of diminishing returns.



To estimate the point of diminishing returns, we use a quadratic polynomial, $y = b_0 + b_1x + b_2x^2$. A very popular women's knit T-shirt was tested for sales appeal and price in six large department stores. In each city, the T-shirts were advertised extensively in the local media, so price and sales initially went up. However, as price increased, sales eventually dropped off. Let x = price per T-shirt in dollars and y = number of T-shirts sold in a day at that price. We have the following data.

City	A	B	C	D	E	F
x	12.97	13.88	15.95	18.50	19.99	22.50
y	23	31	33	29	25	17

To construct our quadratic polynomial, we use multilinear regression with the following table of data values.

$x_1 = x$	12.97	13.88	15.95	18.50	19.99	22.50
$x_2 = x^2$	168.22	192.65	254.40	342.25	399.60	506.25
y	23	31	33	29	25	17

Computer software gives us coefficients for the model $y = b_0 + b_1x_1 + b_2x_2 = b_0 + b_1x + b_2x^2$, which becomes $y = -93.80 + 15.10x - 0.45x^2$. The coefficient of determination is $r^2 = 0.88$ (not too bad!). The curvilinear regression equation $y = -93.80 + 15.10x - 0.45x^2$ is a quadratic curve that opens downward. A little extra mathematics shows that the top point on the curve (point of diminishing returns) occurs when the cost per shirt of $x = \$16.78$ with $y = 32.87$ shirts sold per day. This suggests the knit t-shirts should be priced at \$16.78 and that about 33 of them will sell per day in a large department store.

Use the Internet, school library, popular magazines, or any other source to collect (x, y) data pairs regarding variables of interest to you. Construct a curvilinear regression model from your data and interpret the results.

TECH NOTES

In the Using Technology section at the end of this chapter, you will find a “mini case study” of seven important variables from the economy of the United States for the years 1976 to 1987. Readers interested in applications of multiple regression and the U.S. economy are referred to this material.



Chapter Review

SUMMARY

This chapter discusses linear regression models and inferences related to these models.

- A scatter diagram of data pairs (x, y) gives a graphical display of the relationship (if any) between x and y data. We are looking for a linear relationship.
- For data pairs (x, y) , x is called the *explanatory variable* and is plotted along the horizontal axis. The *response variable* y is plotted along the vertical axis.
- The Pearson product-moment *correlation coefficient* r gives a numerical measurement assessing the strength of a linear relationship between x and y . It is based on a random sample of (x, y) data pairs.
- The value of r ranges from -1 to 1 , with 1 indicating perfect positive linear correlation, -1 indicating perfect negative linear correlation, and 0 indicating no linear correlation.
- If the scatter diagram and sample correlation coefficient r indicate a linear relationship between x and y values of the data pairs, we use the least-squares criteria to develop the equation of the least-squares line

$$\hat{y} = a + bx$$

where \hat{y} is the value of y predicted by the least-squares line for a given x value, a is the y intercept, and b is the slope.

- Methods of testing the population correlation coefficient ρ show whether or not the sample

statistic r is significant. We test the null hypothesis $H_0: \rho = 0$ against a suitable alternate hypothesis ($\rho > 0$, $\rho < 0$, or $\rho \neq 0$).

- Methods of testing the population slope β show whether or not the sample slope b is significant. We test the null hypothesis $H_0: \beta = 0$ against a suitable alternate hypothesis ($\beta > 0$, $\beta < 0$, or $\beta \neq 0$).
- Confidence intervals for β give us a range of values for β based on the sample statistic b and specified confidence level c .
- Confidence intervals for the predicted value of y give us a range of values for y for a specific x value. The interval is based on the sample prediction \hat{y} and confidence level c .
- The *coefficient of determination* r^2 is a value that measures the proportion of variation in y explained by the least-squares line, the linear regression model, and the variation in the explanatory variable x .
- The difference $y - \hat{y}$ between the y value in the data pair (x, y) and the corresponding predicted value \hat{y} for the same x is called the *residual*.
- The *standard error of estimate* S_e is a measure of data spread about the least-squares line. It is based on the residuals.
- Techniques of multiple regression (with computer assistance) help us analyze a linear relation involving several variables.

IMPORTANT WORDS & SYMBOLS

Section 9.1

Paired data values 502
 Scatter diagram 502
 Explanatory variable 502
 Response variable 502
 No linear correlation 503
 Perfect linear correlation 504
 Positive correlation 504
 Negative correlation 505
 Sample correlation coefficient r 506
 Extrapolation 513
 Causation 513
 Lurking variable 513

Section 9.2

Least-squares criterion 521
 Least-squares line $\hat{y} = a + bx$ 522
 Slope b 522
 Intercept a 522
 Meaning of slope 524
 Marginal change 525
 Influential point 525
 Residual 525
 Interpolation 526
 Extrapolation 526
 Coefficient of determination r^2 530

Explained variation 531
 Unexplained variation 531
 Residual plot 537

Section 9.3

Population correlation coefficient ρ 541
 Standard error of estimate S_e 544

Population slope β 547
 Confidence prediction band 549

Section 9.4

Multiple regression 559
 Coefficient of multiple determination 565
 Curvilinear (polynomial) regression 573

VIEWPOINT

Living Arrangements

Male, female, married, single, living alone, living with friends or relatives—all these categories are of interest to the U.S. Census Bureau. In addition to these categories, there are others, such as age, income, and health needs. How strongly correlated are these variables? Can we use one or more of these variables to predict the others? How good is such a prediction? Methods of this chapter can help you answer such questions. For more information regarding such data, visit the Brase/Brase statistics site at <http://www.cengage.com/statistics/brase> and find the link to Census Bureau.

CHAPTER REVIEW PROBLEMS

- Statistical Literacy** Suppose the scatter diagram of a random sample of data pairs (x, y) shows no linear relationship between x and y . Do you expect the value of the sample correlation coefficient r to be close to 1, -1 , or 0?
- Statistical Literacy** What does it mean to say that the sample correlation coefficient r is significant?
- Statistical Literacy** When using the least-squares line for prediction, are results usually more reliable for extrapolation or interpolation?
- Statistical Literacy** Suppose that for $x = 3$, the predicted value is $\hat{y} = 6$. The data pair $(3, 8)$ is part of the sample data. What is the value of the residual for $x = 3$?

In Problems 5–10, parts (a)–(e) involve scatter diagrams, least-squares lines, correlation coefficients with coefficients of determination, tests of ρ , and predictions. Parts (f)–(i) involve standard error of estimate, confidence intervals for predictions, tests of β , and confidence intervals for β .

When solving problems involving the standard error of estimate, testing of the correlation coefficient, or testing of β or confidence intervals for β , make the assumption that x and y are normally distributed random variables. Answers may vary slightly due to rounding.

- Desert Ecology: Wildlife** Bighorn sheep are beautiful wild animals found throughout the western United States. Data for this problem are based on information taken from *The Desert Bighorn*, edited by Monson and Sumner (University of Arizona Press). Let x be the age of a bighorn sheep (in years), and let y be the mortality rate (percent that die) for this age group. For example, $x = 1, y = 14$ means that 14% of the bighorn sheep between 1 and 2 years old die. A random sample of Arizona bighorn sheep gave the following information:

x	1	2	3	4	5
y	14	18.9	14.4	19.6	20.0

$$\Sigma x = 15; \Sigma y = 86.9; \Sigma x^2 = 55; \Sigma y^2 = 1544.73; \Sigma xy = 273.4$$

- Draw a scatter diagram.
- Find the equation of the least-squares line.
- Find r . Find the coefficient of determination r^2 . Explain what these measures mean in the context of the problem.

- (d) Test the claim that the population correlation coefficient is positive at the 1% level of significance.
- (e) Given the lack of significance of r , is it practical to find estimates of y for a given x value based on the least-squares line model? Explain.
6. **Sociology: Job Changes** A sociologist is interested in the relation between x = number of job changes and y = annual salary (in thousands of dollars) for people living in the Nashville area. A random sample of 10 people employed in Nashville provided the following information:

x (Number of job changes)	4	7	5	6	1	5	9	10	10	3
y (Salary in \$1000)	33	37	34	32	32	38	43	37	40	33

- $\Sigma x = 60$; $\Sigma y = 359$; $\Sigma x^2 = 442$; $\Sigma y^2 = 13,013$; $\Sigma xy = 2231$
- (a) Draw a scatter diagram for the data.
- (b) Find \bar{x} , \bar{y} , b , and the equation of the least-squares line. Plot the line on the scatter diagram of part (a).
- (c) Find the sample correlation coefficient r and the coefficient of determination. What percentage of variation in y is explained by the least-squares model?
- (d) Test the claim that the population correlation coefficient ρ is positive at the 5% level of significance.
- (e) If someone had $x = 2$ job changes, what does the least-squares line predict for y , the annual salary?
- (f) Verify that $S_e \approx 2.56$.
- (g) Find a 90% confidence interval for the annual salary of an individual with $x = 2$ job changes.
- (h) Test the claim that the slope β of the population least-squares line is positive at the 5% level of significance.
- (i) Find a 90% confidence interval for β and interpret its meaning.
7. **Medical: Fat Babies** Modern medical practice tells us not to encourage babies to become too fat. Is there a positive correlation between the weight x of a 1-year-old baby and the weight y of the mature adult (30 years old)? A random sample of medical files produced the following information for 14 females:

x (lb)	21	25	23	24	20	15	25	21	17	24	26	22	18	19
y (lb)	125	125	120	125	130	120	145	130	130	130	130	140	110	115

- $\Sigma x = 300$; $\Sigma y = 1775$; $\Sigma x^2 = 6572$; $\Sigma y^2 = 226,125$; $\Sigma xy = 38,220$
- (a) Draw a scatter diagram for the data.
- (b) Find \bar{x} , \bar{y} , b , and the equation of the least-squares line. Plot the line on the scatter diagram of part (a).
- (c) Find the sample correlation coefficient r and the coefficient of determination. What percentage of the variation in y is explained by the least-squares model?
- (d) Test the claim that the population correlation coefficient ρ is positive at the 1% level of significance.
- (e) If a female baby weighs 20 pounds at 1 year, what do you predict she will weigh at 30 years of age?
- (f) Verify that $S_e \approx 8.38$.
- (g) Find a 95% confidence interval for weight at age 30 of a female who weighed 20 pounds at 1 year of age.
- (h) Test the claim that the slope β of the population least-squares line is positive at the 1% level of significance.
- (i) Find an 80% confidence interval for β and interpret its meaning.

8. **Sales: Insurance** Dorothy Kelly sells life insurance for the Prudence Insurance Company. She sells insurance by making visits to her clients' homes. Dorothy believes that the number of sales should depend, to some degree, on the number of visits made. For the past several years, she has kept careful records of the number of visits (x) she makes each week and the number of people (y) who buy insurance that week. For a random sample of 15 such weeks, the x and y values follow:

x	11	19	16	13	28	5	20	14	22	7	15	29	8	25	16
y	3	11	8	5	8	2	5	6	8	3	5	10	6	10	7

$$\Sigma x = 248; \Sigma y = 97; \Sigma x^2 = 4856; \Sigma y^2 = 731; \Sigma xy = 1825$$

- Draw a scatter diagram for the data.
 - Find \bar{x} , \bar{y} , b , and the equation of the least-squares line. Plot the line on the scatter diagram of part (a).
 - Find the sample correlation coefficient r and the coefficient of determination. What percentage of the variation in y is explained by the least-squares model?
 - Test the claim that the population correlation coefficient ρ is positive at the 1% level of significance.
 - In a week during which Dorothy makes 18 visits, how many people do you predict will buy insurance from her?
 - Verify that $S_e \approx 1.731$.
 - Find a 95% confidence interval for the number of sales Dorothy would make in a week during which she made 18 visits.
 - Test the claim that the slope β of the population least-squares line is positive at the 1% level of significance.
 - Find an 80% confidence interval for β and interpret its meaning.
9. **Marketing: Coupons** Each box of Healthy Crunch breakfast cereal contains a coupon entitling you to a free package of garden seeds. At the Healthy Crunch home office, they use the weight of incoming mail to determine how many of their employees are to be assigned to collecting coupons and mailing out seed packages on a given day. (Healthy Crunch has a policy of answering all its mail on the day it is received.)

Let x = weight of incoming mail and y = number of employees required to process the mail in one working day. A random sample of 8 days gave the following data:

x (lb)	11	20	16	6	12	18	23	25
y (Number of employees)	6	10	9	5	8	14	13	16

$$\Sigma x = 131; \Sigma y = 81; \Sigma x^2 = 2435; \Sigma y^2 = 927; \Sigma xy = 1487$$

- Draw a scatter diagram for the data.
- Find \bar{x} , \bar{y} , b , and the equation of the least-squares line. Plot the line on the scatter diagram of part (a).
- Find the sample correlation coefficient r and the coefficient of determination. What percentage of the variation in y is explained by the least-squares model?
- Test the claim that the population correlation coefficient ρ is positive at the 1% level of significance.
- If Healthy Crunch receives 15 pounds of mail, how many employees should be assigned mail duty that day?
- Verify that $S_e \approx 1.726$.
- Find a 95% confidence interval for the number of employees required to process mail for 15 pounds of mail.

- (h) Test the claim that the slope β of the population least-squares line is positive at the 1% level of significance.
- (i) Find an 80% confidence interval for β and interpret its meaning.
10. **Focus Problem: Changing Population and Crime Rate** Let x be a random variable representing percentage change in neighborhood population in the past few years, and let y be a random variable representing crime rate (crimes per 1000 population). A random sample of six Denver neighborhoods gave the following information (Source: *Neighborhood Facts*, The Piton Foundation).

x	29	2	11	17	7	6
y	173	35	132	127	69	53

$$\Sigma x = 72; \Sigma y = 589; \Sigma x^2 = 1340; \Sigma y^2 = 72,277; \Sigma xy = 9499$$

- (a) Draw a scatter diagram for the data.
- (b) Find \bar{x} , \bar{y} , b , and the equation of the least-squares line. Plot the line on the scatter diagram of part (a).
- (c) Find the sample correlation coefficient r and the coefficient of determination. What percentage of the variation in y is explained by the least-squares model?
- (d) Test the claim that the population correlation coefficient ρ is not zero at the 1% level of significance.
- (e) For a neighborhood with $x = 12\%$ change in population in the past few years, predict the change in the crime rate (per 1000 residents).
- (f) Verify that $S_e \approx 22.5908$.
- (g) Find an 80% confidence interval for the change in crime rate when the percentage change in population is $x = 12\%$.
- (h) Test the claim that the slope β of the population least-squares line is not zero at the 1% level of significance.
- (i) Find an 80% confidence interval for β and interpret its meaning.

DATA HIGHLIGHTS: GROUP PROJECTS

Break into small groups and discuss the following topics. Organize a brief outline in which you summarize the main points of your group discussion.

Scatter diagrams! Are they really useful? Scatter diagrams give a first impression of a data relationship and help us assess whether a linear relation provides a reasonable model for the data. In addition, we can spot *influential points*. A data point with an extreme x value can heavily influence the position of the least-squares line. In this project, we look at data sets with an influential point.

x	1	4	5	9	10	15
y	3	7	6	10	12	4

- (a) Compute r and b , the slope of the least-squares line. Find the equation of the least-squares line, and sketch the line on the scatter diagram.
- (b) Notice the point boxed in blue in Figure 9-24. Does it seem to lie away from the linear pattern determined by the other points? The coordinates of that point are (15, 4). Is it an influential point? Remove that point from the model and recompute r , b , and the equation of the least-squares line. Sketch this least-squares line on the diagram. How does the removal of the influential point affect the values of r and b and the position of the least-squares line?
- (c) Consider the scatter diagram of Figure 9-25. Is there an influential point? If you remove the influential point, will the slope of the new least-squares line be larger or smaller than the slope of the line from the original data? Will the correlation coefficient be larger or smaller?

FIGURE 9-24

Scatter Diagram

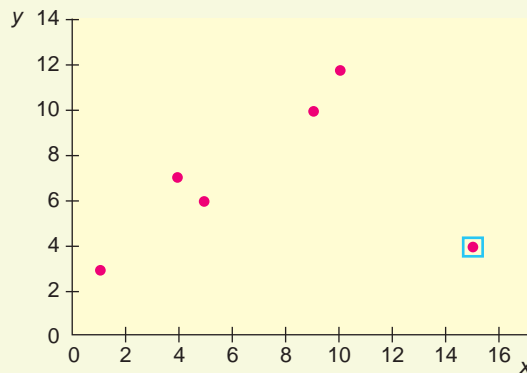
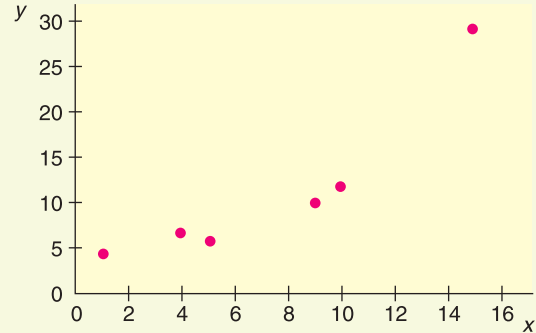


FIGURE 9-25

Scatter Diagram



LINKING CONCEPTS: WRITING PROJECTS

Discuss each of the following topics in class or review the topics on your own. Then write a brief but complete essay in which you summarize the main points. Please include formulas and graphs as appropriate.

1. What do we mean when we say that two variables have a strong positive (or negative) linear correlation? What would a scatter diagram for these variables look like? Is it possible that two variables could be strongly related somehow but have a low *linear* correlation? Explain and draw a scatter diagram to demonstrate your point.
2. What do we mean by the least-squares criterion? Give a very general description of how the least-squares criterion is involved in the construction of the least-squares line. Why do we say the least-squares line is the “best-fitting” line for the data set?
3. In this chapter, we discussed three measures for “goodness of fit” of the least-squares line for given data. These measures were standard error of estimate, correlation coefficient, and coefficient of determination. Discuss the ways in which these measurements are different and the ways in which they are similar to each other. Be sure to include a discussion of explained variation, unexplained variation, and total variation in your answer. Draw a sketch and include appropriate formulas.
4. Look at the formula for confidence bounds for least-squares predictions. Which of the following conditions do you think will result in a *shorter* confidence interval for a prediction?
 - (a) Larger or smaller values for the standard error of estimate
 - (b) Larger or smaller number of data pairs
 - (c) A value of x near \bar{x} or a value of x far away from \bar{x}
 Why would a shorter confidence interval for a prediction be more desirable than a longer interval?
5. If you did not cover Section 9.4, Multiple Regression, omit this problem.

For many applications in statistics, more data lead to more accurate results. In multiple regression, we have more variables (and data) than we have in most simple regression problems. Why will this usually lead to more accurate predictions? Will additional variables *always* lead to more accurate predictions? Explain your answer. Discuss the coefficient of multiple determination and its meaning in the context of multiple regression. How do we know if an explanatory variable has a statistically significant influence on the response variable? What do we mean by a regression model?

6. Use the Internet or go to the library and find a magazine or journal article in your field of major interest to which the content of this chapter could be applied. List the variables used, method of data collection, and general type of information and conclusions drawn.

USING TECHNOLOGY

Simple Linear Regression (One Explanatory Variable)

Application 1

The data in this section are taken from this source:

Based on King, Cuchlaine A. M. *Physical Geography*. Oxford: Basil Blackwell, 1980, pp. 77–86, 196–206.

Throughout the world, natural ocean beaches are beautiful sights to see. If you have visited natural beaches, you may have noticed that when the gradient or dropoff is steep, the grains of sand tend to be larger. In fact, a man-made beach with the “wrong” size granules of sand tends to be washed away and eventually replaced when the proper size grain is selected by the action of the ocean and the gradient of the bottom. Since man-made beaches are expensive, grain size is an important consideration.

In the data that follow, x = median diameter (in millimeters) of granules of sand, and y = gradient of beach slope in degrees on natural ocean beaches.

x	y
0.17	0.63
0.19	0.70
0.22	0.82
0.235	0.88
0.235	1.15
0.30	1.50
0.35	4.40
0.42	7.30
0.85	11.30

1. Find the sample mean and standard deviation for x and y .
2. Make a scatter plot. Would you expect a moderately high correlation and a good fit for the least-squares line?
3. Find the equation of the least-squares line, and graph the line on the scatter plot.
4. Find the sample correlation coefficient r and the coefficient of determination r^2 . Is r significant at the 1% level of significance (two-tailed test)?
5. Test that $\beta > 0$ at the 1% level of significance. Find the standard error of estimate S_e and form an 80% confidence interval for β . As the diameter of granules of sand changes by 0.10 mm, by how much does the gradient of beach slope change?
6. Suppose you have a truckload of sifted sand in which the median size of granules is 0.38 mm. If you want to put this sand on a beach and you don't want the sand to wash away, then what does the least-squares line predict for the angle of the beach? *Note:* Heavy storms that produce abnormal waves may also wash out the sand. However, in the long run, the size of sand granules that remain on the beach or that are brought back to the beach by long-term wave action are determined to a large extent by the angle at which the beach drops off. What range of angles should the beach have if we want to be 90% confident that we are matching the size of our sand granules (0.38 mm) to the proper angle of the beach?
7. Suppose we now have a truckload of sifted sand in which the median size of the granules is 0.45 mm. Repeat Problem 6.

Technology Hints (Simple Regression)

TI-84Plus/TI-83Plus/TI-nspire (with TI-84Plus keypad)

Be sure to set **DiagnosticOn** (under **Catalog**).

- Scatter diagram: Use **STAT PLOT**, select the first type, use **ZOOM** option **9:ZoomStat**.
- Least-squares line and r : Use **STAT**, **CALC**, option **8:LinReg(a + bx)**.
- Graph least-squares line and predict: Press **Y=**. Then, under **VARS**, select **5:Statistics**, then select **EQ**, and finally select item **1:RegEQ**. Press enter. This sequence of steps will automatically set $Y_1 =$ your regression equation. Press **GRAPH**. To find a predicted value, when the graph is showing press the **CALC** key and select item **1:Value**. Enter the x value, and the corresponding y value will appear.
- Testing ρ and β , value for S_e : Use **STAT**, **TEST**, option **E:LinRegTTest**. The value of S_e is in the display as s .
- Confidence intervals for β or predictions: Use formulas from Section 9.3.

Excel 2007

- Scatter plot, least-squares line, r^2 : On home screen, click **Insert** tab. In the **Charts** group, select **Scatter** and choose the first type. Once plot is displayed, *right* click on any data point. Select **trend line**. Under options, check display line and display r^2 .
- Prediction: Use **insert function** (f_x) **> Statistical > Forecast**.
- Coefficient r : Use (f_x) **> Statistical > Correl**.

- Testing β and confidence intervals for b : Use menu selection **Tools > Data Analysis > Regression**.
- Confidence interval for prediction: Use formulas from Section 9.3.

Minitab

- Scatter plot, least-squares line, r^2 , S_e : Use menu selection **Stat > Regression > Fitted line plot**. The value of S_e is displayed as the value of s .
- Coefficient r : Use menu selection **Stat > Basic Statistics > Correlation**.
- Testing β , predictions, confidence interval for predictions: Use menu selection **Stat > Regression > Regression**.
- Confidence interval for β : Use formulas from Section 9.3.

SPSS

SPSS offers several options for finding the correlation coefficient r and the equation of the least-squares line. First enter the data in the data editor and label the variables appropriately in the variable view window. Use the menu choices **Analyze > Regression > Linear** and select dependent and independent variables. The output includes the correlation coefficient, the standard error of estimate, the constant, and the coefficient of the dependent variable with corresponding t values and P -values for two-tailed tests. The display shows the results for the data in this chapter's Focus Problem regarding crime rate and percentage change in population.

SPSS Display

Model Summary				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.927 ^a	.859	.823	22.59076

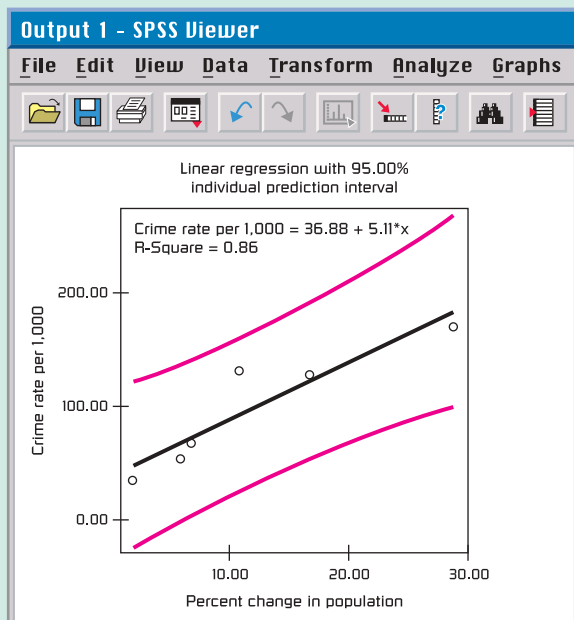
a. Predictors: (Constant), % change in population

Coefficients ^a						
Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	36.881	15.474		2.383	.076
	% change in population	5.107	1.035	.927	4.932	.008

a. Dependent Variable: Crime rate per 1,000

With the menu choices **Graph** ► **Legacy Dialogues** ► **Interactive** ► **Scatterplot**, SPSS produces a scatter diagram with the least-squares line, least-squares equation, coefficient of determination r^2 , and optional prediction bands. In the dialogue box, move the dependent variable to the box along the vertical axis and the independent variable to the box along the horizontal axis. Click the “fit” tab, highlight Regression as method, and check the box to include the constant in the equation. For optional prediction band, check individual, enter the confidence level, and check total. The following display shows a scatter diagram for the data in this chapter’s Focus Problem regarding crime rate and percentage change in population.

SPSS Display for Focus Problem



Multiple Regression

Application 2

Data values in the following study are taken from *Statistical Abstract of the United States*, U.S. Department of Commerce, 103rd and 109th Editions (see Table 9-16). All data values represent annual averages as determined by the U.S. Department of Commerce.

- Construct a regression model with
 - Response variable: x_3 (foreign investments)
 - Explanatory variables: x_5 (GNP), x_6 (U.S. dollar), and x_7 (consumer credit)

What is the coefficient of multiple determination?

- Use a 1% level of significance and test each coefficient for significance (two-tailed test).
- Examine the coefficients of the regression equation. Then explain why you think the following statement is true or false: "If the purchasing power of the U.S. dollar did not change and the GNP did not change, then an increase in consumer credit would likely be accompanied by a reduction in foreign investments."
- Suppose $x_5 = 3500$, $x_6 = 0.975$, and $x_7 = 450$. Predict the level of foreign investment. Find a 90% confidence interval for your prediction.

TABLE 9-16 Economic Data, 1976–1987 (on the data disk)

Year	x_1	x_2	x_3	x_4	x_5	x_6	x_7
1976	10.9	7.61	31	974.9	1718	1.757	234.4
1977	12.0	7.42	35	894.6	1918	1.649	263.8
1978	12.5	8.41	42	820.2	2164	1.532	308.3
1979	17.7	9.44	54	844.4	2418	1.380	347.5
1980	28.1	11.46	83	891.4	2732	1.215	349.4
1981	35.6	13.91	109	932.9	3053	1.098	366.6
1982	31.8	13.00	125	884.4	3166	1.035	381.1
1983	29.0	11.11	137	1190.3	3406	1.000	430.4
1984	28.6	12.44	165	1178.5	3772	0.961	511.8
1985	26.8	10.62	185	1328.2	4015	0.928	592.4
1986	14.6	7.68	209	1792.8	4240	0.913	646.1
1987	17.9	8.38	244	2276.0	4527	0.880	685.5

We will use the following notation:

- x_1 = price of a barrel of crude oil, in dollars per barrel
- x_2 = percent interest on 10-year U.S. Treasury notes
- x_3 = total foreign investments in U.S., in billions of dollars
- x_4 = Dow Jones Industrial Average (DJIA)
- x_5 = Gross National Product, GNP, in billions of dollars
- x_6 = purchasing power of U.S. dollar with base 1983 corresponding to \$1.000
- x_7 = consumer credit (i.e., consumer debt), in billions of dollars

2. Construct a new regression model with
 Response variable: x_4 (DJIA)
 Explanatory variables: x_3 (foreign investments), x_5 (GNP), and x_7 (consumer credit)

What is the coefficient of multiple determination?

- Use a 5% level of significance and test each coefficient for significance (two-tailed test).
- Examine the coefficients of the regression equation; then explain why you think the following statement is true or false: “If the GNP and consumer credit didn’t change but foreign investments increased, the DJIA would likely show a strong increase.”
- Suppose $x_3 = 210$, $x_5 = 4260$, and $x_7 = 650$. Predict the DJIA and find an 85% confidence interval for your prediction.

3. Construct a new regression model with
 Response variable: x_7 (consumer credit)
 Explanatory variables: x_3 (foreign investments), x_5 (GNP), and x_6 (U.S. dollar)

What is the coefficient of multiple determination?

- Use a 1% level of significance and test each coefficient for significance (two-tailed test).
- Examine the coefficients of the regression equation; then explain why you think each of the following statements is true or false: “If both GNP and purchasing power of the U.S. dollar didn’t change, then an increase in foreign

investments would likely be accompanied by a reduction in consumer credit.” “If both foreign investments and purchasing power of the U.S. dollar remained fixed, then an increase in GNP would likely be accompanied by an increase in consumer credit.”

- Suppose $x_3 = 88$, $x_5 = 2750$, and $x_6 = 1.250$. Predict consumer credit and find an 80% confidence interval for your prediction.

Technology Hints (Multiple Regression)

TI-84Plus/TI-83Plus/TI-nspire

Does not support multiple regression.

Excel 2007

On the home screen, click the **Data** tab. Select **Data Analysis** from the Analysis group. In the dialogue box, select **Regression**. On the spreadsheet, the columns containing the explanatory variables need to be adjacent.

Minitab

Use the menu selection **Stat** ► **Regression** ► **Regression**.

SPSS

Use the menu selection **Analyze** ► **Regression** ► **Linear** and select dependent and independent variables.



Cumulative Review Problems

CHAPTERS 7–9

In Problems 1–6, please use the following steps (i) through (v) for all hypothesis tests.

- (i) What is the level of significance? State the null and alternate hypotheses.
- (ii) **Check Requirements** What sampling distribution will you use? What assumptions are you making? What is the value of the sample test statistic?
- (iii) Find (or estimate) the P -value. Sketch the sampling distribution and show the area corresponding to the P -value.
- (iv) Based on your answers in parts (i) to (iii), will you reject or fail to reject the null hypothesis? Are the data statistically significant at level α ?
- (v) **Interpret** your conclusion in the context of the application.

Note: For degrees of freedom $d.f.$ not in the Student's t table, use the closest $d.f.$ that is *smaller*. In some situations, this choice of $d.f.$ may increase the P -value a small amount and thereby produce a slightly more “conservative” answer.

1. **Testing and Estimating μ , σ Known** Let x be a random variable that represents micrograms of lead per liter of water (ug/l). An industrial plant discharges water into a creek. The Environmental Protection Agency has studied the discharged water and found x to have a normal distribution, with $\sigma = 0.7$ ug/l (Reference: *EPA Wetlands Case Studies*).
 - (a) The industrial plant says that the population mean value of x is $\mu = 2.0$ ug/l. However, a random sample of $n = 10$ water samples showed that $\bar{x} = 2.56$ ug/l. Does this indicate that the lead concentration population mean is higher than the industrial plant claims? Use $\alpha = 1\%$.
 - (b) Find a 95% confidence interval for μ using the sample data and the EPA value for σ .

- (c) How large a sample should be taken to be 95% confident that the sample mean \bar{x} is within a margin of error $E = 0.2$ ug/l of the population mean?

2. **Testing and Estimating μ , σ Unknown** Carboxyhemoglobin is formed when hemoglobin is exposed to carbon monoxide. Heavy smokers tend to have a high percentage of carboxyhemoglobin in their blood (Reference: *Laboratory and Diagnostic Tests* by F. Fishbach). Let x be a random variable representing percentage of carboxyhemoglobin in the blood. For a person who is a regular heavy smoker, x has a distribution that is approximately normal. A random sample of $n = 12$ blood tests given to a heavy smoker gave the following results (percent carboxyhemoglobin in the blood).

9.1	9.5	10.2	9.8	11.3	12.2
11.6	10.3	8.9	9.7	13.4	9.9

- (a) Use a calculator to verify that $\bar{x} \approx 10.49$ and $s \approx 1.36$.
 - (b) A long-term population mean $\mu = 10\%$ is considered a health risk. However, a long-term population mean above 10% is considered a clinical alert that the person may be asymptomatic. Do the data indicate that the population mean percentage is higher than 10% for this patient? Use $\alpha = 0.05$.
 - (c) Use the given data to find a 99% confidence interval for μ for this patient.
3. **Testing and Estimating a Proportion p** Although older Americans are most afraid of crime, it is young people who are more likely to be the actual victims of crime. It seems that older people are more cautious about the people with whom they associate. A national survey showed that 10% of all people ages 16–19 have been victims of crime (Reference: *Bureau*

of Justice Statistics). At Jefferson High School, a random sample of $n = 68$ students (ages 16–19) showed that $r = 10$ had been victims of a crime.

- Do these data indicate that the population proportion of students in this school (ages 16–19) who have been victims of a crime is different (either way) from the national rate for this age group? Use $\alpha = 0.05$. Do you think the conditions $np > 5$ and $nq > 5$ are satisfied in this setting? Why is this important?
- Find a 90% confidence interval for the proportion of students in this school (ages 16–19) who have been victims of a crime.
- How large a sample size should be used to be 95% sure that the sample proportion \hat{p} is within a margin of error $E = 0.05$ of the population proportion of all students in this school (ages 16–19) who have been victims of a crime? *Hint:* Use sample data \hat{p} as a preliminary estimate for p .

- Testing Paired Differences** Phosphorous is a chemical that is found in many household cleaning products. Unfortunately, phosphorous also finds its way into surface water, where it can harm fish, plants, and other wildlife. Two methods of phosphorous reduction are being studied. At a random sample of 7 locations, both methods were used and the total phosphorous reduction (mg/l) was recorded (Reference: *Environmental Protection Agency Case Study 832-R-93-005*).

Site	1	2	3	4	5	6	7
Method I:	0.013	0.030	0.015	0.055	0.007	0.002	0.010
Method II:	0.014	0.058	0.017	0.039	0.017	0.001	0.013

Do these data indicate a difference (either way) in the average reduction of phosphorous between the two methods? Use $\alpha = 0.05$.

- Testing and Estimating $\mu_1 - \mu_2$, σ_1 and σ_2 Unknown**

In the airline business, “on-time” flight arrival is important for connecting flights and general customer satisfaction. Is there a difference between summer and winter average on-time flight arrivals? Let x_1 be a random variable that represents percentage of on-time arrivals at major airports in the summer. Let x_2 be a random variable that represents percentage of on-time arrivals at major airports in the winter. A random sample of $n_1 = 16$ major airports showed that $\bar{x}_1 = 74.8\%$, with $s_1 = 5.2\%$. A random sample of $n_2 = 18$ major airports showed that $\bar{x}_2 = 70.1\%$, with $s_2 = 8.6\%$ (Reference: *Statistical Abstract of the United States*).

- Does this information indicate a difference (either way) in the population mean percentage of on-time arrivals for summer compared to winter? Use $\alpha = 0.05$.
- Find a 95% confidence interval for $\mu_1 - \mu_2$.
- What assumptions about the original populations have you made for the methods used?

- Testing and Estimating a Difference of Proportions $p_1 - p_2$**

How often do you go out dancing? This question was asked by a professional survey group on behalf of the National Arts Survey. A random sample of $n_1 = 95$ single men showed that $r_1 = 23$ went out



dancing occasionally. Another random sample of $n_2 = 92$ single women showed that $r_2 = 19$ went out dancing occasionally.

- (a) Do these data indicate that the proportion of single men who go out dancing occasionally is higher than the proportion of single women who do so? Use a 5% level of significance. List the assumptions you made in solving this problem. Do you think these assumptions are realistic?
 - (b) Compute a 90% confidence interval for the population difference of proportions $p_1 - p_2$ of single men and single women who occasionally go out dancing.
7. **Essay and Project** In Chapters 7 and 8 you studied, estimation and hypothesis testing.
- (a) Write a brief essay in which you discuss using information from samples to infer information about populations. Be sure to include methods of estimation and hypothesis testing in your discussion. What two sampling distributions are used in estimation and hypothesis testing of population means, proportions, paired differences, differences of means, and differences of proportions? What are the criteria for determining the appropriate sampling distribution? What is the level of significance of a test? What is the P -value? How is the P -value related to the alternate hypothesis? How is the null hypothesis related to the sample test statistic? Explain.
 - (b) Suppose you want to study the length of time devoted to commercial breaks for two different types of television programs. Identify the types of programs you want to study (e.g., sitcoms, sports events, movies, news, children's programs, etc.). Write a brief outline for your study. Consider whether you will use paired data (such as same time slot on two different channels) or independent samples. Discuss how to

obtain random samples. How large should the sample be for a specified margin of error? Describe the protocol you will follow to measure the times of the commercial breaks. Determine whether you are going to compare the average time devoted to commercials or the proportion of time devoted to commercials. What assumptions will you make regarding population distributions? What graphics might be appropriate? What methods of estimation will you use? What methods of testing will you use?

8. **Critical Thinking** Explain hypothesis testing to a friend, using the following scenario as a model. Describe the hypotheses, the sample statistic, the P -value, the meanings of type I and type II errors, and the level of significance. Discuss the significance of the results. Formulas are not required.

A team of research doctors designed a new knee surgery technique utilizing much smaller incisions than the standard method. They believe recovery times are shorter when the new method is used. Under the old method, the average recovery time for full use of the knee is 4.5 months. A random sample of 38 surgeries using the new method showed the average recovery time to be 3.6 months, with sample standard deviation of 1.7 months. The P -value for the test is 0.0011. The research team states that the results are statistically significant at the 1% level of significance.

9. **Linear Regression: Blood Glucose** Let x be a random variable that represents blood glucose level after a 12-hour fast. Let y be a random variable representing blood glucose level 1 hour after drinking sugar water (after the 12-hour fast). Units are in mg/10 ml. A random sample of eight adults gave the following information (Reference: *American Journal of Clinical Nutrition*, Vol. 19, pp. 345–351).

Jacob Halaska/Photolibrary/Getty Images



$$\Sigma x = 63.8; \Sigma x^2 = 521.56; \Sigma y = 90.7; \\ \Sigma y^2 = 1070.87; \Sigma xy = 739.65$$

x	6.2	8.4	7.0	7.5	8.1	6.9	10.0	9.7
y	9.8	10.7	10.3	11.9	14.2	7.0	14.6	12.2

- Draw a scatter diagram for the data.
- Find the equation of the least-squares line and graph it on the scatter diagram.
- Find the sample correlation coefficient r and the sample coefficient of determination r^2 . Explain the meaning of r^2 in the context of the application.
- If $x = 9.0$, use the least-squares line to predict y . Find an 80% confidence interval for your prediction.
- Use level of significance 1% and test the claim that the population correlation coefficient ρ is not zero. Interpret the results.
- Find an 85% confidence interval for the slope β of the population-based least-squares line. Explain its meaning in the context of the application.